Escola Tècnica Superior d'Enginyeria Informàtica

Universitat Politècnica de València

Estudio de los vuelos nacionales de EEUU

Proyecto de la asignatura

Proyecto II

Grado en Ciencia de Datos

Autores:
Joaquin Carrión Gil
Iván García Donderis
Marcos Gómez Soler
Carles Navarro Esteve
Gonzalo Hurtado Sanhermelando

Tutoras:
Sara Blanc Clavero
María José Ramirez Quintana

Curso 2023 - 2024

ÍNDICE

1. Alcance del proyecto

Nuestro proyecto se llama Estudio de los vuelos nacionales de EEUU. Hemos realizado diversos análisis sobre datos que trataban de los vuelos nacionales de los Estados Unidos. Nos hemos centrado en sacar información relevante para nuestro estudio que la plasmamos en nuestros objetivos.

Principalmente, hemos estudiado el retraso que podían tener los distintos vuelos dependiendo de compañías o aeropuertos, aunque también hemos estudiado la densidad de tráfico aéreo según la zona o incluso el día, distinguiendo entre días festivos y días que no lo son y la distribución de vuelos y aeropuertos en Estados Unidos.

1.1 Objetivos del proyecto

En este proyecto, nos hemos planteado 5 objetivos, los cuales son los siguientes:

1. Estudiar los aeropuertos per cápita, por superficie; analizar la distribución y el tráfico aéreo de los estados de EEUU:

   1.1. Generar una visualización donde se muestre la distribución de aeropuertos per cápita y superficie, y analizar la distribución de los aeropuertos.

   1.2. Analizar el tráfico aéreo de cada estado y ver cómo se distribuyen los estados según el número de vuelos.

   1.3. Ver las relaciones que hay entre las variables de la base de datos y ver qué estados tienen perfiles similares o si hay estados que tienen relación.

2. Realizar un estudio de compañías aéreas que trabajan en Estados Unidos:

   2.1 Correlación entre el número de vuelos de cada compañía y el número de aeropuertos donde trabaja

   2.2 Ranking de mejores y peores aerolíneas según el retraso

   2.3 Correlación entre retraso_salida y retraso_llegada. Esto se hace para comprobar si en el caso de que el avión salga con retraso, si el vuelo se realiza de manera más rápida con el fin de contrarrestar el retraso ocasionado.

   2.4 Correlación entre probabilidad de retraso y número de vuelos

   2.5 Estudio del número de vuelos y retraso para diferentes compañías a lo largo del tiempo

   2.6 Estudio de las causas de retraso para cada compañía seleccionada

   2.7 Visualización geográfica para cada compañía para ver donde trabaja con más frecuencia

   2.8 Estudio de las mejores compañías para cada región de Estados Unidos

3. Estudio de retraso proporcionado por aeropuertos y rutas.

   3.1 Estudio de aeropuertos cuando actúan como origen.

      3.1.1 Detección y estudio de aeropuertos anómalos.

3.1.2 Correlaciones: número total de vuelos vs. probabilidad de retraso y probabilidad de retraso vs. media en minutos de vuelos retrasados.

3.2 Estudio de aeropuertos cuando actúan como destino.

3.2.1 Detección y estudio de aeropuertos anómalos.

3.2.2 Correlaciones: número total de vuelos vs. probabilidad de retraso y probabilidad de retraso vs. media en minutos de vuelos retrasados.

3.3 Creación índices de rendimiento para cada aeropuerto para cuando actúa como origen y destino. Correlación entre ambos índices.

3.4 Clustering de los aeropuertos.

3.5 Estudio de las rutas.

3.5.1 Detección de rutas anómalas.

3.5.2 Estudio de rutas de interés.

3.5.3 Estudio de si la compañía influye de manera estadísticamente significativa en el rendimiento de una ruta.

3.5.3.1 ANOVA.

3.5.3.2 Modelo de regresión lineal.

4. Número de vuelos y retrasos en fechas significativas en comparación a sus vísperas y otros días

4.1 Cantidad media de vuelos los días de la semana de los 3 años estudiados

4.2 Cantidad de vuelos por meses de los 3 años estudiados

4.3 Cantidad de vuelos los 4 días festivos escogidos y sus días cercanos

4.4 Retraso medio de los vuelos por meses los 3 años estudiados

4.5 Retraso medio de los vuelos por días de la semana de los 3 años estudiados

4.6 Comparativa del retraso los 4 días festivos y sus días cercanos con las medias del día de la semana (de ese mes) y del mes correspondiente.

5. Realizar una aplicación web, que muestre información relevante de los vuelos entre dos aeropuertos a elegir por el usuario.

1.2. Utilidad del estudio

En cuanto a la utilidad del estudio, creemos que este estudio podría ser útil para aquellas personas que decidan realizar un viaje entre estados de los Estados Unidos de América, para que obtengan información de diferentes variables interesantes a tener en cuenta cuando vayan a realizar el viaje. También podría interesarles a las diferentes compañías aéreas o aeropuertos de los Estados Unidos de América para intentar tomar decisiones que mejoren sus circunstancias de cara a los pasajeros. Además, ya que vamos a analizar la distribución de los aeropuertos por estados, puede ayudar a empresas privadas o al propio gobierno de EEUU a construir nuevos aeropuertos en estados donde la distribución no sea óptima.

Por otro lado, pensamos que puede ser novedoso ya que a la hora de escoger el vuelo para ir a algún sitio determinado, los usuarios no poseen de la información necesaria como para poder decidir entre distintas compañías o vuelos en función del retraso es más, solamente se conocen opciones que ayudan a comparar el precio de los vuelos. Nosotros pensamos que esta información no es suficiente, ya que un vuelo que sea barato puede acarrear un retraso elevado, por lo que igual es mejor opción que el vuelo sea un poco más caro pero puntual. No hemos visto ningún proyecto o página web parecida antes, así que creemos que los resultados son de gran ayuda para el usuario, ya que el tiempo es oro.

2. Configuración del proyecto

2.1. Fuentes de datos

Aparte de las mencionadas en el hito 1, hemos actualizado este hito, cuyos cambios son los siguientes **(véase en el anexo 1).** También hemos añadido dos fuentes de datos:

- Una que contiene el área en kilómetros cuadrados, la población, el número de aeropuertos y el número de vuelos de todos los estados de EEUU, esta base de datos se utilizará para realizar el objetivo 1, donde analizaremos la distribución de aeropuertos de los estados.
- Una que contiene las coordenadas de los aeropuertos de Estados Unidos, para así poder representar las localizaciones de los aeropuertos deseados por el usuario en un mapa.

2.2. Integración y transformación de los datos

Hemos actualizado el hito 2 **(véase en el anexo 2)**. En este hito, nos encargamos de la principal parte de integración y transformación de los datos. A partir de aquí creamos un nuevo dataframe limpiado para que todos podamos trabajar con él. Sin embargo, durante el desarrollo de los objetivos seguimos integrando nuevos datos y creando nuevas variables derivadas para obtener los resultados buscados, donde está explicada su obtención dentro de los anexos de cada objetivo.

3. Resultados obtenidos

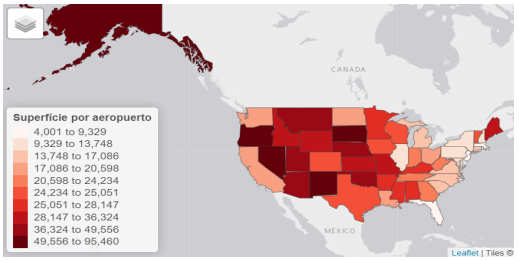
3.1 Analizar la distribución de los aeropuertos y el tráfico aéreo de los estados de USA

Para ampliar información de este objetivo (más gráficas y más conclusiones) y ver el funcionamiento del código, **véase en el anexo 3: Objetivo_1.**

Para estudiar la distribución de aeropuertos por estado creímos conveniente centrarnos en analizar los aeropuertos per cápita y los aeropuertos por superficie.

Para ello, primero realizamos un gráfico para ver el número de aeropuertos por estado y así poder ver la cantidad de aeropuertos en cada zona. El estado con más aeropuertos fue Texas con 28 aeropuertos y el que menos cantidad de aeropuertos tiene fue Delaware con cero aeropuertos, este último no lo tendremos en cuenta para el análisis al no contar con aeropuertos en su área.

Para analizar los estados respecto a los aeropuertos por superficie, hicimos un mapa de calor que nos permitió ver con más claridad qué zonas tienen un ratio superficie/aeropuerto mayor. Se puede ver en el mapa que los estados que se sitúan al oeste del país suelen tener más superficie por aeropuerto, es decir, que tienen menos aeropuertos en proporción a su superficie. También se puede apreciar cómo estos estados con un ratio de superficie por aeropuerto alto, son estados con una superficie mayor, comprobaremos si existe esta correlación más adelante en el objetivo.



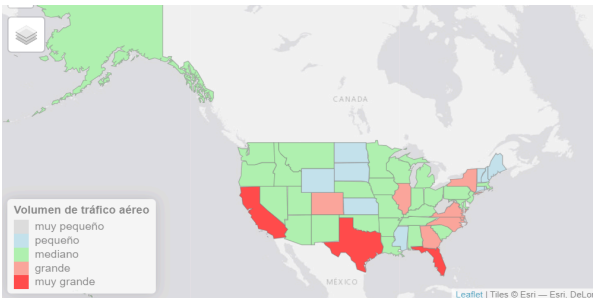
Una vez analizada la relación entre la superficie y los aeropuertos pasamos a analizar el ratio de población y aeropuertos. Para ello realizamos un gráfico que muestra los estados con mayor número de habitantes por aeropuerto. Los resultados fueron que el estado con mayor ratio habitantes/aeropuerto es Connecticut con un valor de 3.626.205 hab/aer. Por otro lado, el estado con menor ratio es el de Alaska con 38609.63 habitantes por aeropuerto.

Para ver la distribución de los estados de EEUU según el número de vuelos realizamos un gráfico de barras a partir de la variable discretizada que clasificaba el volumen de vuelos de cada estado.

Observamos como los estados siguen una distribución normal, donde la mayoría de los estados tienen un tamaño de vuelos mediano. Los tamaños de vuelos pequeños y grandes tienen una frecuencia muy similar de aproximadamente 9 estados en cada uno. Y después están los tamaños de vuelos muy pequeños y muy grandes que tienen una frecuencia muy inferior a los otros.

Para analizar el tráfico aéreo, hicimos un mapa de calor para ver la información de una forma más visual. Para ello utilizamos el número de vuelos que ha habido por estado en 2023.

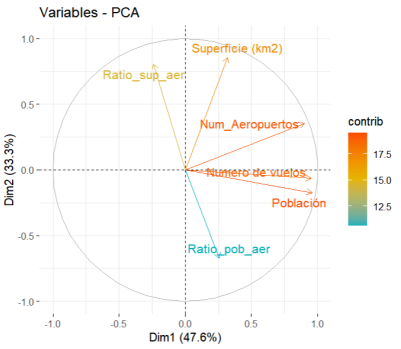
Se puede observar como los estados costeros son los que más densidad de vuelos tienen mientras que los estados que se ubican en el interior del país tienen menos vuelos. Esto se puede deber a que en los estados costeros existe un mayor turismo y por tanto hay una mayor afluencia de gente. También es de considerar que estos estados actúan como estados puente, es decir, que reciben un gran número de vuelos



internacionales y después se desplazan a cualquier parte del país mediante un vuelo nacional desde el propio aeropuerto del estado. Aun así existen estados que no cumplen con esta regla, tenemos el ejemplo de Colorado que se encuentra en el interior del país y aun así tiene mucha densidad de vuelos, también observamos como Maine a pesar de ser un estado costero no cuenta con una gran cantidad de vuelos.

Para finalizar con el objetivo decidimos realizar un PCA sobre nuestra base de datos. Primero estudiamos el número de componentes principales necesarios. Para ello, empleamos el criterio del codo y vimos que con las dos primeras PCs explicamos un poco más del 80% de la variabilidad de los datos.

Aplicando el PCA con dos componentes principales, realizamos el gráfico de loadings para ver las correlaciones que hay entre las variables y ver sus contribuciones a las diferentes componentes.



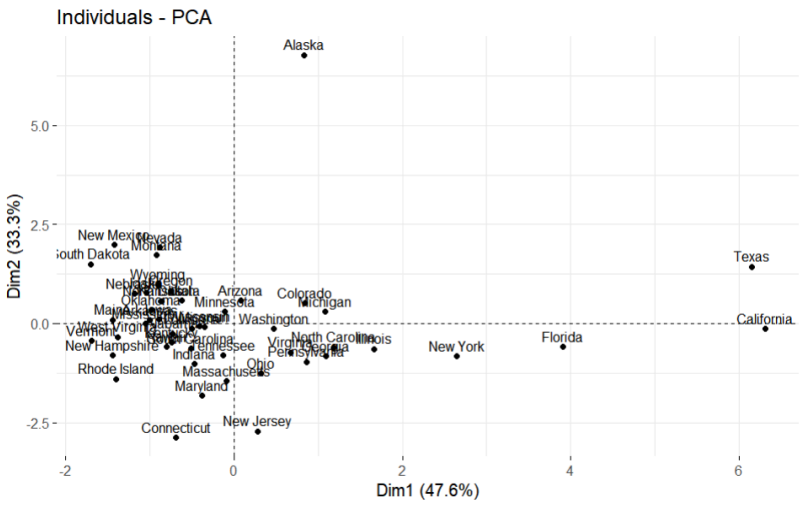
Las variables que más contribuyen en la PC1 son las variables “Población”, “num_aeropuertos” y “numero_de_vuelos” que están correlacionadas positivamente entre ellas, es decir, si el valor en una de estas variables es alto, en el de las otras variables correlacionadas también será alto.

En cambio para la PC2 las variables que más contribuyen son “Superficie”, “Ratio_sup_aer” y “ratio_pob_aer”. “Superficie” y “Ratio_sup_aer” están correlacionadas positivamente entre sí, pero estas dos están correlacionadas negativamente con “ratio_pob_aer”.

Para ver el comportamiento de los estados de EEUU y ver las relaciones que hay entre ellos representamos el gráfico de scores del PCA, gracias al cual de pudimos ver que la mayoría de estados tienen un perfil muy similar para todas las variables de nuestra base de datos, pero hay algún estado que está alejado de esta nube de puntos.

Texas y California tienen valores parecidos tanto en las variables de la primera componente como en los valores de la segunda componente y ambos estados tienen valores muy altos en cuanto a la población, el número de aeropuertos y el número de vuelos por eso están alejados de la nube de puntos.

En cambio, Alaska, dista de la nube de puntos ya que tiene valores muy altos en cuanto a la superficie y a los ratios de población por aeropuertos y superficie por aeropuertos. Comparando Alaska con New Jersey podemos decir que tienen valores similares para las variables que contribuyen más a la primera componente, pero para las variables que contribuyen más a la segunda componente Alaska tendrá unos valores más altos que New Jersey.



3.2 Estudio de las compañías aéreas

Para afrontar este objetivo, nos planteamos diversas ideas que pudieran ser interesantes para el proyecto. Para ampliar información (más gráficas y más conclusiones) y ver el funcionamiento del código, **véase en el anexo 4: Objetivo_2.**

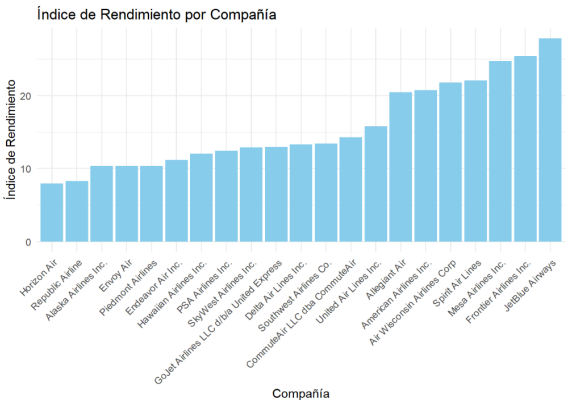
Para empezar, analizamos las compañías con mayor y menor número de vuelos en Estados Unidos: La compañía con mayor número de vuelos y con una gran diferencia del resto fue Southwest Airlines Co., superando el millón de vuelos en 2023. Otras compañías destacables fueron Delta Air Lines Inc. y American Airlines Inc.

Por otra parte, la compañía con menor número de vuelos en 2023 fue GoJet Airlines LLC, siendo unas 30 veces menor que Southwest Airlines Co. Otra que podemos destacar entre las cinco con menores vuelos es Hawaiian Airlines Inc., una compañía que tendremos en cuenta para nuestro estudio, y tiene tan pocos porque solo trabaja en Hawaii y en algunos escasos estados de Estados Unidos.

Quisimos analizar también el número de aeropuertos donde operan las aerolíneas. Pensábamos que cuantos más aeropuertos trabajan, mayor número de vuelos deberían tener. El resultado fue SkyWest Airlines Inc. como la compañía en la que más aeropuertos trabaja, y vimos que Southwest Airlines Co., la compañía con más números de vuelos, no entraba en este top5. Es por eso que decidimos hacer el test de Pearson para ver si existía correlación entre el número de aeropuertos y el número de vuelos por compañía, y el resultado fue que no existía relación entre estas dos variables.

Después de esto, nos centramos en el estudio del retraso de las compañías. Creamos un nuevo dataframe con nuevas variables interesantes como el total de salidas, media retraso, probabilidad de retraso y índice de rendimiento (variable calculada a partir de la media de retraso y de la probabilidad del retraso), todas agrupadas por el nombre de la compañía.

Según la distribución del índice de rendimiento, podemos clasificar las compañías en Buena, Normal y Mala. Gracias a esto, pudimos sacar una lista de las mejores y peores compañías aéreas de Estados Unidos.



Mejores compañías por orden en función del retraso: Horizon Air, Republic Airline y Envoy Air.

Peores compañías por orden en función del retraso: JetBlue Airways, Frontier Airlines Inc. y Mesa Airlines Inc.

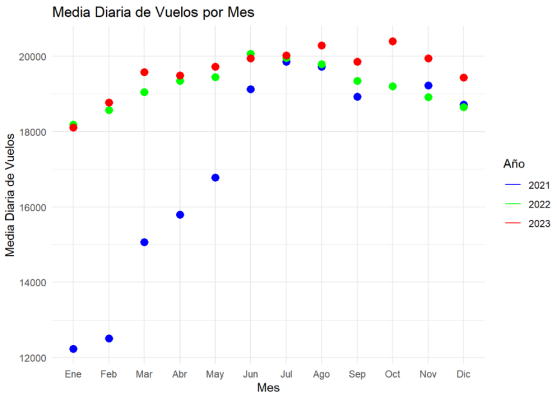
Después de sacar la lista de mejores y peores compañías, pensamos en si existía correlación entre el número de vuelos de la compañía y el índice de rendimiento que habíamos sacado. Haciendo el test de correlación de Pearson concluimos que no existe relación entre estas dos variables.

Otra idea que se nos ocurrió es que siempre se ha dicho que hay compañías, por lo menos aquí en Europa, que aceleran su vuelo si ha salido con retraso para que al final el retraso total no sea tan significativo. Quisimos demostrar si esto era verídico en Estados Unidos, e hicimos un gráfico y un test de correlación de Pearson para investigarlo. Para ello comparamos la media de retraso de salida y la media de retraso de llegada para cada compañía, y pudimos observar que no existía ninguna gran diferencia entre estas dos variables, y según el test, que las dos variables estaban fuertemente relacionadas. Por tanto la conclusión es que no es verídico que en Estados Unidos aceleren el viaje para paliar los retrasos en la salida, según nuestros datos.

Después de esto, empezamos con el análisis de las compañías seleccionadas. Para hacerlo, elegimos 5 compañías, las cuales son: Southwest Airlines Co.,Delta Air Lines Inc., Hawaiian Airlines Inc., JetBlue Airways y Republic Airline. En primer lugar, estudiamos la media diaria de vuelos por mes para los años 2021, 2022 y 2023 para todas las compañías, para ver cual era el comportamiento.

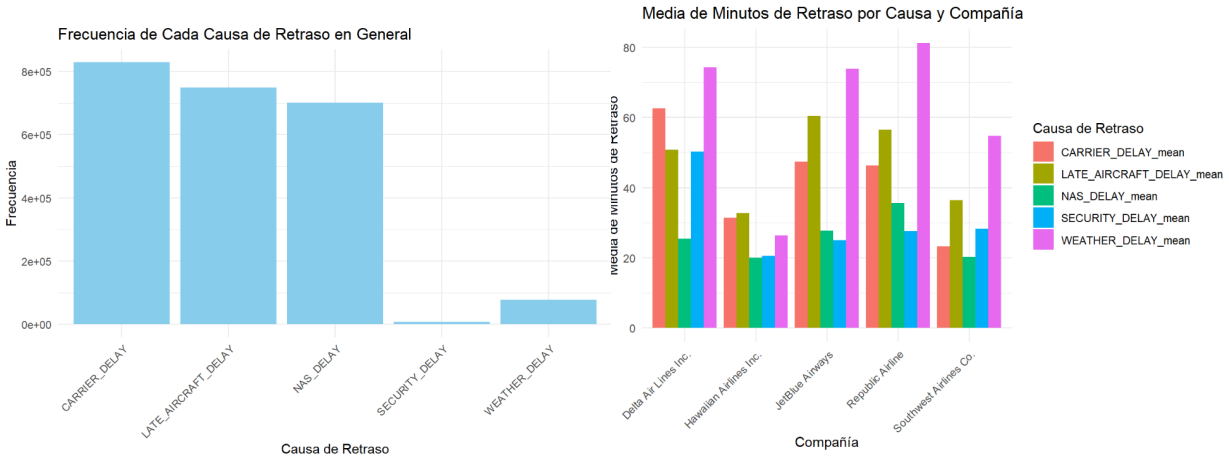
Analizando el gráfico pudimos sacar diversas conclusiones:

- El año 2021 empezó con muy pocos vuelos debido a la crisis del Covid-19, pero aumentó significativamente durante el año para llegar a la normalidad.
- En general, se puede ver que la tendencia para todos los años es la misma, aumentando en los meses de verano y disminuyendo en los meses de invierno.



También estudiamos la frecuencia de las causas de retraso para todas las compañías y la media de retraso por causa, donde sacamos las siguientes conclusiones:

- Los tres motivos más frecuentes son el retraso de la compañía, el retraso que lleva el avión y el retraso del sistema nacional de aviación.
- La peor causa que se puede experimentar es por motivos metereológicos,aunque sea poco frecuente, es la que más media de retraso en minutos tiene.



<u>Southwest Airlines Co.:</u> Se trata de la compañía con mayor número de vuelos y con la mayor probabilidad de retraso con un 52% de sus vuelos. Se observa que sigue la misma distribución de número de vuelos que la mayoría, aumentando en verano, pero destacamos el pico más alto en octubre de 2023, donde investigando, encontramos que esta aerolínea compró más aviones en esas fechas.

El número de vuelos retrasados aumenta en verano, pero desciende bastante después de verano, siendo inversamente proporcional al número de vuelos, por lo que confirmamos una mejora en cuanto a la probabilidad de retraso en los últimos meses del año.

La causa de retraso más frecuente es la llegada tardía del avión.

Trabaja mayoritariamente en estados del sur y suroeste.

Delta Air Lines Inc.: Sigue la distribución general, subiendo bastante el número de vuelos en los meses de verano y después bajando en invierno. La distribución de vuelos retrasados es más rara. Tienen un pico alto en junio y julio, pero tienen una gran bajada en los últimos meses del año, demostrando que ellos también reducen su probabilidad de retraso al final del año 2023. Según el índice de retraso, es normal.

La causa de retraso más frecuente es el retraso de la compañía.

Trabaja mayoritariamente en estados del sur y noreste.

Hawaiian Airlines Inc.: Se trata de una compañía mucho más pequeña porque, como comentamos anteriormente, solo trabaja con aeropuertos de Hawaii y sus correspondientes destinos de Estados Unidos. Aumenta el número de vuelos en verano y disminuye en invierno. Lo que es destacable es que el número de vuelos retrasados tiene el pico en abril, y baja bastante en verano, lo que implica que la compañía mejora su servicio en esta estación, cuando sabe que tiene más trabajo, al ser Hawaii un destino turístico en verano.

La causa de retraso más frecuente es el retraso de la compañía.

Trabaja mayoritariamente en Hawaii.

JetBlue Airways: Tiene un comportamiento diferente a los demás, al bajar el número de vuelos en los meses de verano. Al investigarlo, nos dimos cuenta que se reducen debido a la falta de personal en el estado de Nueva York, que es uno de los estados donde más opera esta aerolínea.

Sobre los vuelos retrasados, destacar que tiene un índice de rendimiento malo, por tanto no actuará muy bien, y que el pico de más vuelos retrasados es en julio, justo el mismo mes con el pico de menor media de número de vuelos. Además, tiene el peor índice de rendimiento

La causa de retraso más frecuente es el retraso de la compañía.

Trabaja mayoritariamente en estados del noreste,sureste y oeste.

Republic Airline: Republic Airline tiene un comportamiento parecido a JetBlue Airways, lo que le hace ser distinto a la mayoría. Baja en los últimos meses del año y en verano.

Tiene un índice de rendimiento bueno, lo podemos destacar también viendo que la causa más frecuente de retraso no es culpa suya (como la mayoría de compañías) sino que es el retraso del Sistema Aéreo Nacional.

Trabaja mayoritariamente en estados del noreste, medio oeste y sureste.

Por último, hemos clasificado los estados en regiones y hemos observado cuáles eran las 5 mejores compañías para cada región y en qué estado era donde funcionaban bien.

En resumen, podemos concluir que Republic Airline es la aerolínea con mejor rendimiento en la mayoría de las regiones, con una fuerte presencia en el Medio Oeste, Medio Atlántico y Sur. SkyWest Airlines Inc. lidera en la región Oeste, mientras que en los territorios, hay una mayor variación de aerolíneas destacadas.

3.3 Estudio de aeropuertos y rutas

En este objetivo llevaremos a cabo un estudio sobre los distintos aeropuertos y rutas presentes en nuestra base de datos teniendo como objeto de estudio principalmente el retraso. Para ampliar información de este objetivo (más gráficas y más conclusiones) y ver el funcionamiento del código, **véase en el anexo 5: Objetivo_3.**

Comenzaremos con el estudio de los aeropuertos. Antes de comenzar a explicar los resultados, es importante conocer que un aeropuerto puede tener dos roles en cualquier vuelo, ser el nodo origen o ser el nodo destino. En ambos escenarios el aeropuerto juega un papel que afecta al correcto funcionamiento de un vuelo, es por ello que estudiaremos a los aeropuertos distinguiendo por la posición en la que actúen en un vuelo.

Empezamos estudiando los aeropuertos cuando actúan como origen. Primero identificamos aeropuertos anómalos, es decir, aeropuertos que poseían medias de retraso en sus vuelos retrasados excesivamente altas. De estos 18 aeropuertos anómalos en el origen, podemos destacar que las regiones predominantes a la que pertenecían era la oeste y la medio oeste con 7 y 6 aeropuertos respectivamente de estos 18.

Quisimos observar también, si estos aeropuertos anómalos en el origen poseían causas de retraso diferentes al resto de aeropuertos. Observamos que tanto los anómalos como el resto de aeropuertos seguían la misma tendencia en los retrasos siendo la causa más frecuente 'CARRIER DELAY', es decir, retraso causado por compañía.

A continuación, mediante diversos test de correlación de Pearson pudimos resolver diversas cuestiones:

- El número total de vuelos de un aeropuerto como origen guarda una relación moderada positiva estadísticamente significativa con la probabilidad de retraso de dicho aeropuerto (coef. de correlación de Pearson = 0.36). Es decir, conforme aumentan los vuelos de un aeropuerto tenderá a aumentar su probabilidad de retraso ligeramente. Esta relación se observaba de una manera más fuerte en los aeropuertos que poseían un número considerable de vuelos. Para los aeropuertos con menos vuelos había una gran variación en la probabilidad de retraso, teniendo aeropuertos con mucha probabilidad de retraso y otros que no proporcionaban apenas vuelos retrasados a pesar de tener los mismos vuelos.
- Por otro lado, la media de retraso en minutos de los vuelos retrasados de un aeropuerto está inversamente correlacionada con la probabilidad de retraso de dicho aeropuerto cuando actúa como origen (coef. de correlación de Pearson = -0.468). Por tanto, a pesar de que haya aeropuertos que proporcionen un mayor número de vuelos retrasados, estos se retrasan menos que otros que poseen una mayor probabilidad de retraso. Esto nos puede llevar a pensar que los aeropuertos con mayor probabilidad de retraso puede que lleven a cabo estrategias que les ayuden a disminuir el tiempo de retraso. Es por eso que, para evaluar la calidad de los aeropuertos en función del retraso, más adelante haremos como hemos hecho antes en el objetivo de compañías, donde hemos calculado un índice de rendimiento basándonos en las variables media de retraso y probabilidad de retraso.

Después, se realizó el mismo estudio para los aeropuertos pero centrándonos en cuando estos actúan como destino. Al igual que antes, encontramos 18 aeropuertos anómalos con medias de retraso extremadamente altas. De los 18 anteriores y de estos 18, 6 aeropuertos son comunes. De estos 6 aeropuertos anómalos comunes podemos destacar que son aeropuertos con poco volumen de vuelos, con probabilidades de retraso similares tanto en el origen como en el destino y siendo estas no muy altas. Volviendo a los aeropuertos anómalos en el destino, ahora la región medio oeste es la más frecuente con 7 aeropuertos seguida de la región sur con 6. Podemos decir que la región medio oeste está presente con fuerza tanto para los aeropuertos que son anómalos en el origen como para los que son anómalos en el destino.

Al igual que sucedía antes, ni los aeropuertos anómalos en el destino, ni los aeropuertos anómalos comunes siguen tendencias diferentes al resto de aeropuertos en lo que a las causas de retraso se refiere, es decir, la causa más frecuente sigue siendo el retraso causado por la compañía.

Volvemos a realizar los mismos test de correlación para los aeropuertos cuando actúan como destino:

- En el caso de la relación entre el número total de vuelos y la probabilidad de retraso, esta sigue siendo estadísticamente significativa y positiva (coef. de correlación de Pearson=0.17). Sin embargo, es más débil que en los aeropuertos de origen, por tanto, a medida que un aeropuerto tiene más vuelos cuando actúa como destino su probabilidad de retraso tiende a subir ligeramente pero con menos intensidad que cuando el aeropuerto actúa como origen.
- Por otro lado, la relación entre la media en minutos de los vuelos retrasados y la probabilidad de retraso sigue siendo significativa. Sin embargo, la correlación es mucho más débil a pesar de seguir estando estas variables inversamente correlacionadas (coef. de correlación de Pearson=-0.109)

Como hemos visto, un aeropuerto puede ser el nodo origen o destino de un vuelo. Para poder resumir el comportamiento de cada aeropuerto en ambas posiciones, generamos un índice de actuación que nos indicará cómo de bien ha actuado un aeropuerto en el origen o en el destino. Este índice se calcula como un producto entre la probabilidad de retraso y la media de retraso de cada aeropuerto, por lo que, cuánto menor sea el índice significa que mejor habrá sido la actuación del aeropuerto ya sea en el origen o en el destino.

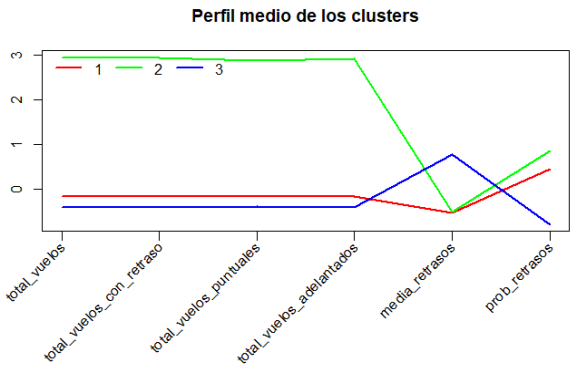
Nos podemos plantear si un aeropuerto que actúa bien también actuará bien en el destino. Para tratar esta cuestión, hemos realizado un nuevo test de correlación donde obtenemos un coeficiente de correlación de Pearson de 0.65 y siendo está correlación estadísticamente significativa. Por tanto, podemos decir que se trata de una relación positiva fuerte entre ambos índices de actuación, es decir, en la mayoría de casos, un aeropuerto que actuó bien como nodo origen, también actuó bien como nodo destino, y viceversa.

A continuación, con el fin de poder agrupar nuestros aeropuertos en función de sus características realizamos un clustering a partir de un resumen general de la actuación de cada aeropuerto. Antes de realizar el clustering prescindimos de los aeropuertos anómalos comunes ya que, podrían interferir en la correcta creación de los clusters y llevar a establecer falsas conclusiones. Tampoco contaremos con la media de los adelantos de cada aeropuerto porque era una variable con poco coeficiente de variación y no aportaba mucha información adicional.

Para realizar el clustering, tras barajar diversos métodos, tanto jerárquicos como de partición, decidimos quedarnos con 3 clusters proporcionados con el algoritmo de k-medias, puesto que presentaba uno de los mejores coeficientes de Silhouette y era el que menos aeropuertos mal clasificados presentaba.

A partir de gráfico del anterior gráfico de líneas podemos ver que cada cluster posee las siguientes características:

- **Cluster 1**: Representa a los aeropuertos con bajo número de vuelos (baja actividad) con una probabilidad de retraso notable. Formado por 185 aeropuertos.
- **Cluster 2**: Formado por aeropuertos con bastantes vuelos (alta actividad) y con una notable probabilidad de retraso, a pesar de ello, estos retrasos no parecen ser muy largos. Este cluster posee 28 aeropuertos.
- **Cluster 3**: Agrupa aeropuertos con bajo número de vuelos (baja actividad) y con



retrasos largos. Sin embargo, poseen una baja probabilidad de retraso. Formado por 140 aeropuertos.

Pasamos ahora al estudio de las rutas. Sabemos que una ruta es el recorrido que sigue un avión en un determinado vuelo, es decir, del aeropuerto A vuela al aeropuerto B, esto constituirá la ruta A-B. Una vez que hemos entendido lo que es una ruta comenzamos con el estudio. Antes de comenzar, decidimos prescindir de las rutas que poseen menos de 100 vuelos. Seguiremos un enfoque similar al de los aeropuertos, primeramente identificamos las rutas anómalas, es decir, rutas que poseen valores extremadamente altos en su retraso medio. El aeropuerto que aparece en más rutas anómalas figura también como el primer aeropuerto presente en un mayor número de rutas, hablamos del aeropuerto DFW. Además, este aeropuerto destaca por ser el cuarto con más vuelos en 2023.

Una vez identificadas las rutas anómalas, observamos si la tendencia en sus causas de retraso difiere del resto de rutas. Concluimos que, al igual que sucedía con los aeropuertos anómalos, tanto las rutas anómalas como las no anómalas tienen como causa de retraso más frecuente el retraso causado por la compañía.

Después, buscamos relaciones entre variables realizando tests de correlación de Pearson donde obtuvimos los siguientes resultados:

- Existe una correlación significativa entre la distancia y la media de retraso de una ruta. Sin embargo, el coeficiente de correlación de Pearson es muy cercano a 0, por lo que no existe correlación entre estas variables.
- Por otro lado, volvemos a encontrar otra correlación significativa pero esta vez entre la distancia y la probabilidad de retraso de una ruta. Esta vez, a partir del coeficiente de correlación de Pearson, podemos decir que se trata de una correlación moderada positiva, es decir, a medida que la distancia a recorrer en la ruta es mayor, la probabilidad de retraso de la ruta tiende a aumentar ligeramente.
- Luego, al contrario que sucedía en los aeropuertos, no encontramos una correlación estadísticamente significativa entre el número total de vuelos de una ruta y su probabilidad de retraso.
- Si estudiamos la correlación entre el total de vuelos de una ruta y su distancia, obtenemos que es estadísticamente significativa y que ambas variables se encuentran inversamente correlacionadas de manera moderada. Podríamos decir que, cuánta mayor distancia entre dos aeropuertos (mayor distancia en esa ruta) el número de vuelos entre ambos tiende a disminuir ligeramente.

A continuación, estudiaremos ciertas rutas en concreto y rutas en las que participan ciertos aeropuertos que consideramos de interés. En primer lugar, realizamos un estudio de la ruta con mayor número de vuelos en 2023, la ruta HNL-OGG. De esta ruta podemos destacar que ambos aeropuertos son de Hawai. Tras estudiar el comportamiento de la misma pudimos observar que en los meses de verano es donde más vuelos tiene, y a su vez, es donde menos vuelos retrasados y donde menos se retrasan sus vuelos en promedio. Este repunte en verano puede estar relacionado con el turismo que recibe Hawai durante estos meses. Podemos destacar también, que el comportamiento de esta ruta guarda una gran relación con el comportamiento que sigue la compañía Hawaiian Airlines, comportamiento explicado en apartados anteriores.

Continuamos centrándonos en las rutas del aeropuerto con más volumen de tráfico en 2023, ATL. Para ello, obtenemos cuáles son las 10 mejores y peores rutas donde participa este aeropuerto empleando en índice de actuación para poder calificar las rutas. Después, estudiamos si los aeropuertos con los que mejor rinde ATL tienen características comunes entre sí. Centrándonos en la región de estos aeropuertos podemos destacar que la región sur es la que está más presente pero no observamos una tendencia predominante.Luego, estudiamos a qué clusters pertenecían estos 10

aeropuertos con los que mejor rendía. Tras ello, pudimos determinar que el aeropuerto de ATL parece rendir mejor con aeropuertos con poco volumen de vuelos ya que, todos los aeropuertos pertenecían al primer o tercer cluster.

Por otro lado, para los aeropuertos con los que peor rendía ATL, podemos destacar que no hay una región que destaque pero llama la atención que rinde mal con 3 aeropuertos de la región medio-atlántica y siendo estos 3 del estado de Florida. En cuánto a los clusters a los que pertenecen, tampoco se observa una tendencia predominante, sin embargo, podemos decir que estos 10 aeropuertos con los que peor rinde se caracterizan por tener probabilidades de retraso notables, características propias de los clusters 1 y 2.

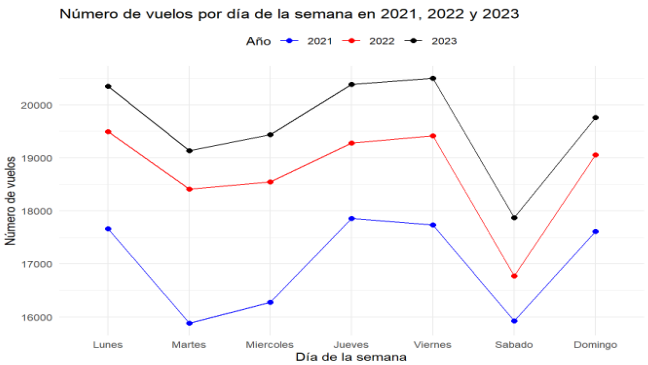
Para finalizar este objetivo, nos centramos en las rutas de los aeropuertos FAI y FLL, siendo el primero un aeropuerto con un buen índice de actuación y el segundo un aeropuerto con mal índice de actuación. Estudiando las rutas de ambos aeropuertos observamos que la compañía parece influir en la forma en la que actúa una determinada ruta. Además, como se ha ido mencionando a lo largo de este objetivo, la causa de retraso mayoritaria es la causada por las compañías.

Por ello, realizamos un ANOVA para determinar si el factor compañía influía de forma significativa en la actuación de una ruta. Tras realizarlo, observamos que el resultado ofrece un nivel de significancia muy alto (p-valor<0.001). Por tanto, podemos concluir que la compañía encargada de la ruta tiene un efecto bastante significativo en en el comportamiento de la misma, es decir, la compañía influye tanto en la probabilidad de retraso de la ruta como en el retraso medio de la misma ya que, el índice de rendimiento se calcula a partir de esas dos variables.

Para finalizar, realizamos un modelo de regresión lineal para explicar el comportamiento de una ruta en función de la compañía. La compañía que elegimos como referencia para el modelo es la compañía Southwest Airlines puesto que es la compañía responsable de más vuelos. A partir de los resultados del modelo podemos concluir que el modelo es estadísticamente significativo. Además, sabemos que las compañías con coeficientes negativos poseerán unos índices de rendimiento menores a la compañía Southwest, es decir, las compañías proporcionarán en sus rutas un mejor rendimiento que Southwest, siempre y cuando sean estadísticamente significativas (p-valor<0.05). Por otro lado, las compañías con coeficientes positivos, proporcionarán en sus rutas un peor rendimiento que Southwest. Por último, el R^2 ajustado del modelo es de 0.3998 lo que representa que el 40% aproximadamente de la variabilidad en el índice de rendimiento de una ruta es explicado por las diferencias entre las compañías que se encargan de la ruta.

3.4 Número de vuelos y retraso en los días festivos y sus vísperas

Para ampliar información de este objetivo(más gráficas y más conclusiones) y ver el funcionamiento del código, **véase en el anexo 6: Objetivo_4.**



Ahora continuamos con el siguiente objetivo, que busca observar si en los festivos y sus días cercanos los vuelos se ven afectados y sea por un aumento o retraso de estos o en la cantidad disponible al público.

En primer lugar, trás realizar Web Scraping para obtener los datos de unos días festivos en Estados Unidos, seleccionamos aquellos que nos parecen más importantes y donde debería notarse más si hay posibles variaciones influenciadas por el día que es. Utilizamos los días de Navidad, Año Nuevo, el Día de la Independencia y Acción de Gracias.

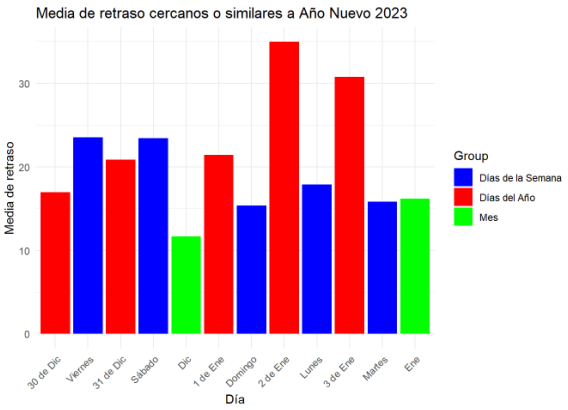
Lo primero que observamos tras realizar una agrupación por días de la semana de los 3 años, es que todos tienen el mismo perfil y solo cambia la cantidad de vuelos, aumentando consecutivamente cada año que pasa.

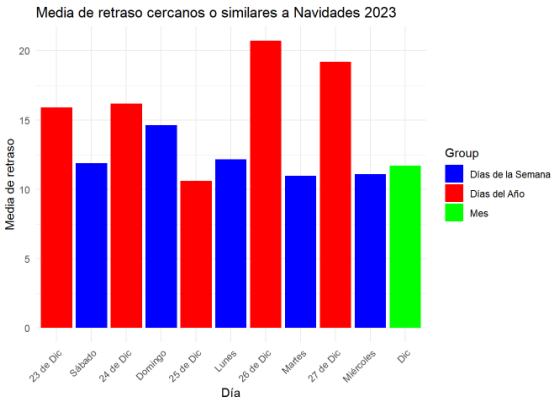
Según lo que vemos en los siguientes gráficos tras observar las cantidades de de vuelos agrupados por días de los meses de los festivos y la línea temporal de días cercanos sabemos que hay una clara reducción de los vuelos las vísperas en comparación con las cifras que debería tener.

Podemos sacar como conclusiones que los días festivos si que tienen una relación directa con la cantidad de vuelos cercanos a esas fechas. El propio día suele tener una bajada de esta cifra, mientras que los días cercanos suelen ser afectados también con una bajada, pero más pequeña. Pero hemos observado un comportamiento un poco diferente en las fechas de Acción de Gracias, considerado que es el festivo más importante, ya que para suministrar la demanda los días previos y posteriores a este suelen aumentar la cantidad de vuelos y el día del festivo hay una gran reducción.

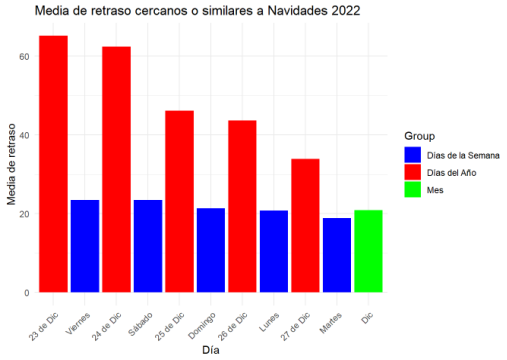
Para continuar hemos realizado análisis similares pero para ver la comparativa con el retraso medio ya sea por días o por meses. Hemos visto que no hay una distribución que se siga muy claramente en ninguno de los dos casos pero en la gráfica por meses vemos que sí que suele ser mayor cercano a verano y en fechas navideñas.

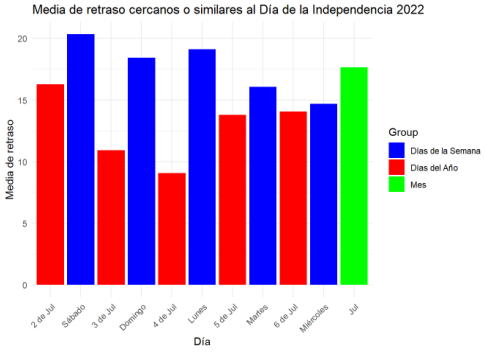
Si vamos más en concreto a las festividades estudiadas anteriormente, no hay claro ningún patrón en el mismo festivo en diferentes años, pero sí que podemos ver algunas similitudes entre los distintos festivos del mismo años. En 2023 hay un aumento del retraso los días posteriores, pese a que en algunos este es muy pequeño y el día del festivo hay un descenso, menos en Año Nuevo que sorpresivamente aumenta.





En 2022, en cambio, los festivos navideños tienen unas cifras de retraso mucho más elevadas los días cercanos y el propio festivo que las medias de los días y del mes, mientras que las otras dos festividades analizadas son justo lo contrario, todos los días de la franja que estamos escogiendo (5 días) tienen un retraso menor.





Y en 2021 si que no vemos ningún tipo de relación ya que en unos festivos son los días previos y en otros los posteriores los que aumentan el retraso, únicamente se mantiene constante que el propio día de la festividad tiene un retraso medio un poco menor, comparado con el día de la semana y el mes.

3.5 Aplicación

Para ampliar información de este objetivo(más gráficas y más conclusiones) y ver el funcionamiento del código, **véase en el anexo: Objetivo_5.**

Finalmente, pasaremos al último objetivo el cual tendrá como finalidad desarrollar una aplicación que dados dos aeropuertos por el usuario, calcule cuál es la mejor ruta además de mostrar información relevante, la cual será calculada a partir de ciertas variables creadas anteriormente, es decir, juntando la información que hemos ido recogiendo hasta ahora, trataremos de recomendar al usuario la mejor ruta con el fin de que experimente el menor retraso posible.

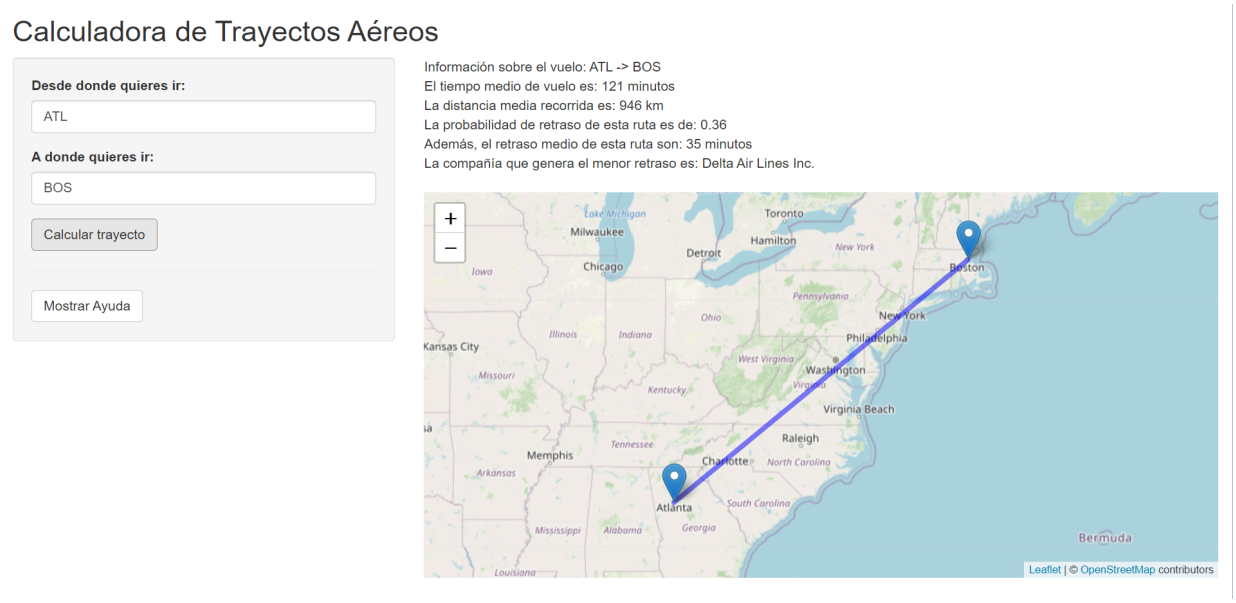
La base de esta aplicación reside en la teoría de grafos. Esto se debe a que para encontrar la ruta más corta, hemos realizado los siguientes pasos: En primer lugar, hemos generado un grafo, el cual hemos modelizado de la siguiente manera: Los vértices son los distintos aeropuertos que tenemos dentro de nuestra base de datos y, entre cada par de vértices, existirá una arista si hay un vuelo directo entre ambos aeropuertos. Dichas aristas tendrán un peso, el cual corresponderá al tiempo del vuelo entre los aeropuertos. En segundo lugar, una vez creado el grafo, aplicamos el algoritmo bfs, el cual busca el camino más corto entre dos vértices teniendo en cuenta el número de aristas que pertenecen al camino. Creemos que esta es la mejor forma de proceder, ya que aunque el tiempo de vuelo en sí pueda llegar a ser más corto, un posible transbordo de más puede elevar el tiempo total de viaje mucho.

A continuación, pasamos al cálculo de los diversos parámetros que constituirán el resultado de la consulta. Dichos parámetros son los siguientes:

- Distancia media recorrida: Se calcula una media de todas las distancias recorridas por todos los vuelos cuyo origen y destino coinciden con los introducidos por el usuario.
- Tiempo medio de vuelo: Se calcula una media de todos los tiempos de vuelo por todos los vuelos cuyo origen y destino coinciden con los introducidos por el usuario.
- Probabilidad de retraso: A partir de la variable ARR_DELAY, se comprueba si el valor para cada uno de los vuelos con el mismo origen y destino que los deseados es positivo. En dicho caso, se suma uno a la variable p_retraso. Una vez hemos recorrido todos los vuelos, se divide entre el total de estos vuelos.
- Retraso medio: Se procede de la misma forma que para la distancia media y el tiempo medio. En este caso no miramos la columna ARR_DELAY, sino que miramos la columna salida_retraso y usamos solo los valores que no son 0.
- Mejor compañía: En este caso, comprobamos para todas las compañías que ofrecen dicho vuelo, cuál es la que tiene un mejor índice de rendimiento (variable explicada en el objetivo 2).

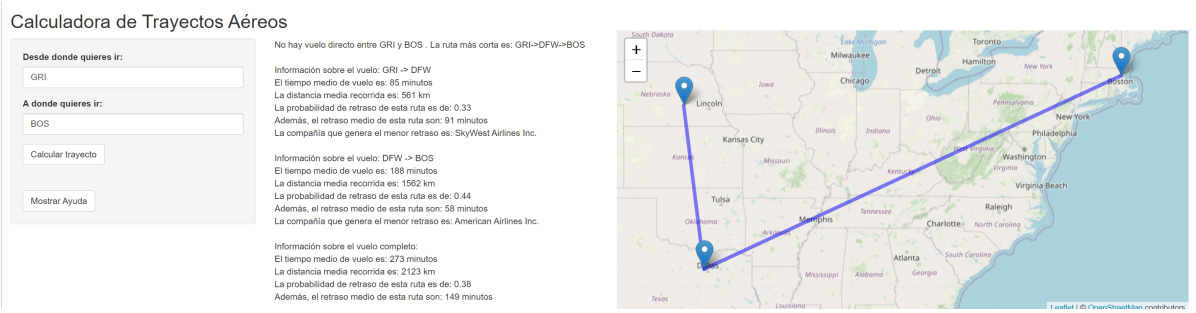
Por otro lado, para que la aplicación funcione correctamente, es necesario que el usuario introduzca los códigos de los aeropuertos. Como no todos saben estos códigos, la aplicación cuenta con una especie de tabla de consulta donde el usuario puede buscar el estado o la ciudad y así, poder introducir correctamente el código.

Por último, pasamos a los resultados. A la hora de calcular la mejor ruta, hay dos posibles situaciones:



En primer lugar, que entre los dos aeropuertos introducidos exista un vuelo directo. En este caso, la aplicación calculará los parámetros correspondientes a dicho vuelo:

Por otro lado, que entre los dos aeropuertos introducidos no exista un vuelo directo. En este caso, la aplicación dividirá el camino más corto calculado en varios vuelos directos, calculará los parámetros de cada uno de los vuelos directos y, finalmente, calculará un resumen total de la ruta completa:



4. Lecciones aprendidas para mis futuros proyectos en Ciencia de Datos

A lo largo de este proyecto, hemos tenido la oportunidad de aprender diferentes lecciones que nos serán muy valiosas de cara a futuros proyectos en Ciencia de Datos.

En primer lugar, hemos aprendido a desenvolvernos en el lenguaje de programación de R, un lenguaje que hasta este año era desconocido para nosotros. Entre otras cosas, hemos aprendido a utilizar distintas librerías como pueden ser dplyr, ggplot2, tmap o shiny, las cuales muy posiblemente tendremos que dominar en un futuro para el análisis de datos.

En segundo lugar, consideramos que hemos mejorado notablemente en el tema del trabajo en equipo y de la organización. El año pasado fue uno de nuestros mayores puntos de mejora y, este año, nos hemos centrado muy seriamente en mejorarlo, de manera que pudiéramos llevar el proyecto al día y no sufrir los días cercanos a la entrega.

Finalmente, en este proyecto hemos hecho uso de distintas técnicas, las cuales hemos tenido la posibilidad de aprender este curso como pueden ser clustering, PCA o teoría de grafos entre otros. Esto, nos ha permitido perfeccionar nuestros conocimientos sobre estos tipos de análisis, los cuales es posible que más adelante formen parte de nuestro día a día.

5. Bibliografía

1. Bureau of Transportation Statistics. (n.d.). *Airline data*. U.S. Department of Transportation. Retrieved May 29, 2024, from https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGK&QO_fu146_anzr=b0-gvzr
2. Telemundo Atlanta. (2023, April 7). *Cuatro grandes aerolíneas reducirán sus vuelos desde y hacia Nueva York durante este verano y esta es la razón*. Telemundo Atlanta. https://www.telemundoatlanta.com/2023/04/07/cuatro-grandes-aerolneas-reduciran-sus-vuelos-desde-y-hacia-nueva-york-durante-este-verano-y-esta-es-la-razn/
3. Aviación Digital. (2023, April 7). *Southwest Airlines incrementa su flota con 108 nuevos aviones Boeing 737 MAX*. Aviación Digital. https://aviaciondigital.com/southwest-airlines-incrementa-su-flota-con-108-nuevos-aviones-boeing-737-max/
4. Datos Mundial. (n.d.). Estados de los EE.UU. Datos Mundial. Retrieved May 29, 2024, from https://www.datosmundial.com/america/usa/estados.php
5. Public Holidays. (2023). Días feriados 2023 in los Estados Unidos - Comparación de los Estados. Public Holidays. Retrieved May 29, 2024, from https://www.public-holidays.us/US_ES_2023_All
6. Miquar. (n.d.). Explore: flights.csv | airports.csv | airlines.csv. Kaggle. Retrieved May 29, 2024, from https://www.kaggle.com/code/miquar/explore-flights-csv-airports-csv-airlines-csv/input?select=airports.csv

6. Anexos

En este apartado, mencionaremos todos los anexos que sirven para tener una visión completa de nuestro proyecto. Contamos con los siguientes:

Anexo 1: Hito 1
Anexo 2: Hito 2
Anexo 3: Objetivo 1
Anexo 4: Objetivo 2
Anexo 5: Objetivo 3
Anexo 6: Objetivo 4
Anexo 7: Objetivo 5

El anexo 1 y 2 estarán presentes en este documento, los demás anexos, los añadiremos en formato HTML para una buena visualización y los adjuntamos a la tarea. En el anexo 2, además de meter la información aquí en la memoría, añadiremos un HTML y los resultados del mismo por si se quiere ver el proceso.

Anexo 1:

1.Alcance

Búsqueda de fuentes
Hemos consultado una gran cantidad de fuentes de datos relativas a vuelos en EEUU. Primero encontramos múltiples bbdd que contaban con mucha información sobre estos, pero eran de pago. Posteriormente otra que hemos barajado albergaba información sobre todos los vuelos a nivel nacional que tuvieron lugar en los Estados Unidos de América. Hemos hallado varías páginas sobre los días festivos en cada estado de América que las usamos haciendo Web Scraping. También hemos hallado páginas donde se detalla la información de los estados (Superfície, población…) la cual usamos haciendo Web Scraping

Criterios seguidos para la selección de las fuentes
El criterio que hemos seguido para la selección de las fuentes de datos es que no fueran de pago y tuvieran suficiente información. Además buscamos que no tuvieran una limpieza de datos muy exhaustiva ya realizada, para poder trabajar con los datos y buscar unos objetivos a alcanzar con estos datos.

2. Técnicas de obtención de datos y extracción

Para la obtención de datos y extracción, hemos utilizado principalmente dos técnicas. En primer lugar, hemos descargado ficheros en .csv de las bbdd relacionadas con los vuelos que posteriormente hemos pasado a un libro de excel. Por otro lado, no siempre hemos tenido la oportunidad de descargar una tabla, por lo que hemos tenido que optar a realizar Web Scraping de páginas relacionadas con los días festivos para tener de esta manera un excel con esta información y para la información de los estados de EEUU.

3. Análisis de las fuentes: interpretación de los datos, valoración de su utilidad en el proyecto

Contamos con **cuatro** fuentes de datos diferentes que hemos escogido tras aplicar nuestros criterios de selección. El eje principal de nuestro proyecto girará en torno a la base de datos de los vuelos a nivel nacional de USA, contamos con la información respectiva a cada vuelo de 2023 sobre el aeropuerto de origen y de destino, la fecha, la hora de salida y de llegada, entre otras variables.

Además hemos añadido dos nuevas bases de datos referente a los años 2021 y 2022. (Conseguimos descargar las bases de datos por meses, así que en el siguiente hito veremos cómo las integramos).

La tercera fuente de datos es una que contiene el área en kilómetros cuadrados, la población, el número de aeropuertos y el número de vuelos de todos los estados de EEUU, esta base de datos se utilizará para realizar el objetivo 1, donde analizaremos la distribución de aeropuertos de los estados.

La última de estas cuatro es la de los días festivos en EEUU. Nos indica los días festivos de cada estado y también cuando es fin de semana.

A partir de la información que manejamos con estas bases de datos podríamos comparar los retrasos en las rutas entre aeropuertos grandes y las rutas entre aeropuertos pequeños así como el retraso medio. Intentaremos determinar la causa del retraso apoyándonos en los datos de la meteorología del lugar de salida y llegada, y comprobar si el retraso en la salida influye en el tiempo de duración del vuelo retrasado, es decir, si el hecho de que salgan con retraso provoca que el avión vaya más rápido. También, podríamos observar si en días emblemáticos (como acción de gracias o el 4 de julio) y en sus vísperas hay un mayor número de vuelos en comparación al resto de días.

Además, tenemos el código de las compañías responsables de cada vuelo, y a partir de esta información, podemos realizar una integración con otra base de datos para identificar las compañías y podremos comparar qué compañías son las que tienen mayores retrasos en los vuelos.

Por otra parte, dejando de lado el tema de los retrasos, podríamos determinar cada cuántos kilómetros hay un aeropuerto en EEUU y en sus respectivos estados (aeropuertos por km^2).

4. Análisis de los campos, formatos y tipo de información de cada fuente y valoración del cruce de datos de distintas fuentes para nuestro proyecto.

Nuestra base de datos de vuelos contiene una gran cantidad de datos, abarcando desde 2021 hasta 2023, con más de 500,000 filas por mes y manteniendo una consistencia elevada, con pocos datos faltantes. Para nuestro estudio detallado, nos enfocaremos en la base de datos de 2023, dado que este año proporciona un volumen considerable de datos. No obstante, utilizaremos los datos de los años anteriores para realizar algunas comparaciones en relación con nuestros objetivos.

Dentro de esta base de datos, disponemos de una amplia variedad de variables de interés, tales como:

- Día del vuelo
- Código de la aerolínea
- Aeropuerto de origen
- Aeropuerto de destino
- Hora de salida prevista
- Hora de salida real
- Hora de llegada prevista
- Hora de llegada real

En nuestras fuentes de datos, hemos tenido que transformar algunas columnas, ya que no estaban en un formato adecuado para los análisis que queremos llevar a cabo. Esta transformación se detalla en el Hito 2.

Además, dependiendo del objetivo específico en el que estemos centrados, podremos crear nuevas bases de datos utilizando la librería *dplyr*. Esta capacidad de generar bases de datos específicas nos

permitirá realizar una amplia gama de nuevos análisis, tales como estudios por compañías aéreas, por aeropuertos, entre otros.

Tenemos un gran potencial con nuestros datos para cruzarlo con las diferentes bases de datos obtenidas y poder sacar un análisis interesante, concluyendo nuestra búsqueda de datos de forma exitosa .

Anexo 2:

2.Interés y alcance del proyecto.

2.1 Explica el objetivo principal de tu proyecto ¿qué presenta este estudio?

El objetivo principal de nuestro proyecto es realizar un estudio general de los vuelos nacionales de Estados Unidos de América de 2023 analizando cómo se comportan en lo referente a retrasos, número de vuelos, compañías, aeropuertos, rutas etc.

2.2 Explica para qué y para quién podría ser de utilidad este estudio

Este estudio podría ser útil para aquellas personas que decidan realizar un viaje entre estados de los Estados Unidos de América, para que obtengan información de diferentes variables interesantes a tener en cuenta cuando vayan a realizar el viaje. También podría interesarles a las diferentes compañías aéreas o aeropuertos de los Estados Unidos de América para intentar tomar decisiones que mejoren sus circunstancias de cara a los pasajeros. Además, ya que vamos a analizar la distribución de los aeropuertos por estados, puede ayudar a empresas privadas o al propio gobierno de EEUU a construir nuevos aeropuertos en estados donde la distribución no sea óptima.

2.3 ¿Por qué piensas que es novedoso? ¿has visto estudios similares?

Pensamos que este estudio puede ser algo novedoso debido a que no conocemos muchas opciones que nos ayuden a tomar decisiones a la hora de viajar con los datos de retrasos de compañías o de aeropuertos, únicamente hay opciones que nos ayudan a comparar los precios de los viajes. También porque mediante este estudio se puede dar a conocer mejor cómo funciona la red de vuelos de los Estados Unidos.

No hemos visto estudios similares ni como mencionamos antes, ninguna página web que ofrezca estos análisis. Pensamos que es importante para el cliente conocer no solo las compañías o aeropuertos con vuelos más baratos, sino saber también el retraso que pueden ocasionar, ya que el tiempo es oro.

2.4 Alcance (objetivos definitivos del proyecto)

Define los objetivos del análisis de datos de tu proyecto. Se deben presentar 5 objetivos de análisis.

Los 5 objetivos que queremos investigar son:

1. Estudiar los aeropuertos per cápita, por superficie; analizar la distribución y el tráfico aéreo de los estados de EEUU:

1.1. Generar una visualización donde se muestre la distribución de aeropuertos per cápita y superficie, y analizar la distribución de los aeropuertos.

1.2. Analizar el tráfico aéreo de cada estado y ver cómo se distribuyen los estados según el número de vuelos.

1.3. Ver las relaciones que hay entre las variables de la base de datos y ver qué estados tienen perfiles similares o si hay estados que tienen relación.

2. Realizar un estudio de compañías aéreas que trabajan en Estados Unidos:

   2.1 Correlación entre el número de vuelos de cada compañía y el número de aeropuertos donde trabaja

   2.2 Ranking de mejores y peores aerolíneas según el retraso

   2.3 Correlación entre retraso_salida y retraso_llegada. Esto se hace para comprobar si en el caso de que el avión salga con retraso, si el vuelo se realiza de manera más rápida con el fin de contrarrestar el retraso ocasionado.

   2.4 Correlación entre probabilidad de retraso y número de vuelos

   2.5 Estudio del número de vuelos y retraso para diferentes compañías a lo largo del tiempo

   2.6 Estudio de las causas de retraso para cada compañía seleccionada

   2.7 Visualización geográfica para cada compañía para ver donde trabaja con más frecuencia

   2.8 Estudio de las mejores compañías para cada región de Estados Unidos

3. Estudio de retraso proporcionado por aeropuertos y rutas.

   3.1 Estudio de aeropuertos cuando actúan como origen.

      3.1.1 Detección y estudio de aeropuertos anómalos.

      3.1.2 Correlaciones: número total de vuelos vs. probabilidad de retraso y probabilidad de retraso vs. media en minutos de vuelos retrasados.

   3.2 Estudio de aeropuertos cuando actúan como destino.

      3.2.1 Detección y estudio de aeropuertos anómalos.

      3.2.2 Correlaciones: número total de vuelos vs. probabilidad de retraso y probabilidad de retraso vs. media en minutos de vuelos retrasados.

   3.3 Creación índices de rendimiento para cada aeropuerto para cuando actúa como origen y destino. Correlación entre ambos índices.

   3.4 Clustering de los aeropuertos.

   3.5 Estudio de las rutas.

      3.5.1 Detección de rutas anómalas.

3.5.2 Estudio de rutas de interés.

3.5.3 Estudio de si la compañía influye de manera estadísticamente significativa en el rendimiento de una ruta.

3.5.3.1 ANOVA.

3.5.3.2 Modelo de regresión lineal.

4. Número de vuelos y retrasos en fechas significativas en comparación a sus vísperas y otros días

4.1 Cantidad media de vuelos los días de la semana de los 3 años estudiados

4.2 Cantidad de vuelos por meses de los 3 años estudiados

4.3 Cantidad de vuelos los 4 días festivos escogidos y sus días cercanos

4.4 Retraso medio de los vuelos por meses los 3 años estudiados

4.5 Retraso medio de los vuelos por días de la semana de los 3 años estudiados

4.6 Comparativa del retraso los 4 días festivos y sus días cercanos con las medias del día de la semana (de ese mes) y del mes correspondiente.

5. Realizar una aplicación web, que muestre información relevante de los vuelos entre dos aeropuertos a elegir por el usuario.

3. Calidad y Análisis exploratorio.

3.1. Integración

En nuestra base de datos principal, teníamos los datos descargados por meses. Nos interesaba trabajar con ellos en un único dataframe, por lo que creamos un código en R para juntar todos en una misma base de datos. Esto lo hemos hecho para los tres años de datos.

Después, a través de la función write.csv pasamos este archivo de datos a formato csv para poder trabajar con él.

```
library(dplyr)
df_ene=read.csv('C:/Users/MAYTE/OneDrive/Escritorio/2º GCD/2 CUATRI/PROY II/VUELOS/vuelos_ene.csv')
df_feb=read.csv('C:/Users/MAYTE/OneDrive/Escritorio/2º GCD/2 CUATRI/PROY II/VUELOS/vuelos_feb.csv')
df_mar=read.csv('C:/Users/MAYTE/OneDrive/Escritorio/2º GCD/2 CUATRI/PROY II/VUELOS/vuelos_mar.csv')
df_abr=read.csv('C:/Users/MAYTE/OneDrive/Escritorio/2º GCD/2 CUATRI/PROY II/VUELOS/vuelos_abr.csv')
df_may=read.csv('C:/Users/MAYTE/OneDrive/Escritorio/2º GCD/2 CUATRI/PROY II/VUELOS/vuelos_may.csv')
df_jun=read.csv('C:/Users/MAYTE/OneDrive/Escritorio/2º GCD/2 CUATRI/PROY II/VUELOS/vuelos_jun.csv')
df_jul=read.csv('C:/Users/MAYTE/OneDrive/Escritorio/2º GCD/2 CUATRI/PROY II/VUELOS/vuelos_jul.csv')
df_ago=read.csv('C:/Users/MAYTE/OneDrive/Escritorio/2º GCD/2 CUATRI/PROY II/VUELOS/vuelos_ago.csv')
df_sep=read.csv('C:/Users/MAYTE/OneDrive/Escritorio/2º GCD/2 CUATRI/PROY II/VUELOS/vuelos_sep.csv')
df_oct=read.csv('C:/Users/MAYTE/OneDrive/Escritorio/2º GCD/2 CUATRI/PROY II/VUELOS/vuelos_oct.csv')
df_nov=read.csv('C:/Users/MAYTE/OneDrive/Escritorio/2º GCD/2 CUATRI/PROY II/VUELOS/vuelos_nov.csv')
df_dec=read.csv('C:/Users/MAYTE/OneDrive/Escritorio/2º GCD/2 CUATRI/PROY II/VUELOS/vuelos_dec.csv')

df_oct=rename(df_oct, MKT_CARRIER=MKT_UNIQUE_CARRIER)
df_oct=rename(df_oct, OP_CARRIER=OP_UNIQUE_CARRIER)
df_nov=rename(df_nov, MKT_CARRIER=MKT_UNIQUE_CARRIER)
df_nov=rename(df_nov, OP_CARRIER=OP_UNIQUE_CARRIER)
df_dec=rename(df_dec, MKT_CARRIER=MKT_UNIQUE_CARRIER)
df_dec=rename(df_dec, OP_CARRIER=OP_UNIQUE_CARRIER)


df_completo=bind_rows(df_ene,df_feb,df_mar,df_abr,df_may,df_jun,df_jul,df_ago,df_sep,df_oct,df_nov,df_dec)

write.csv(df_completo, file='C:/Users/MAYTE/OneDrive/Escritorio/2º GCD/2 CUATRI/PROY II/VUELOS/vuelos_completo.csv')
```

Imagen con el código de integración.

Además de la base de datos de vuelos, para el objetivo 1 hemos utilizado una tabla para la cual hemos tenido que utilizar técnicas de Web Scraping para obtener la información de la superficie, de la población y del nombre del estado. A continuación, hemos sacado el número de vuelos por estado y el número de aeropuerto por estado utilizando la base de datos principal mediante el uso de funciones que permitiera contar las filas de la base de datos por aeropuerto (y así saber el número de vuelos por estado) y también contar los aeropuertos por estado.

3.2. Limpieza y transformación de los datos

Antes de nada, lo que hemos hecho es describir todas las variables que tenemos en nuestra base de datos y hemos eliminado aquellas variables que nos parecían irrelevantes para nuestro proyecto.

Más adelante, hemos detectado los valores faltantes que teníamos en nuestros datos. La gran mayoría de los datos faltantes, tras buscar una causa, nos hemos dado cuenta de que provienen de vuelos cancelados y desviados, los cuales no nos interesan, por lo que podemos eliminarlos.

A continuación, hemos realizado una serie de transformaciones en nuestros datos. Concretamente, lo que haremos es cambiar los datos relativos a fechas al formato 'hh:mm'.

Llegados a este punto, lo que vamos a hacer es crear nuevas características/variables, las cuales nos serán útiles en un futuro. Lo que vamos a hacer es que a partir de cada una de las columnas relativas al retraso (DEP_DELAY y ARR_DELAY) calcularemos dos columnas derivadas las cuales serán: salida_adelanto, salida_retraso, llegada_adelanto y llegada_retraso. Si el valor en la columna DEP_DELAY es positivo, significa que ha habido retraso y se pondrá el valor en salida_retraso, mientras que si es negativo, significa que hay adelanto y se pondrá en salida_adelanto. Emplearemos el mismo procedimiento para ARR_DELAY. Por otro lado, a nuestra base de datos limpiada, le añadiremos una nueva columna la cual nos indica la compañía con la cual se ha realizado cada uno de los vuelos que tenemos en nuestros datos.

Finalmente, en cuanto a la tabla de número de aeropuertos, superfície, población, etc hemos realizado una transformación de los datos que consiste, en primer lugar, eliminar los puntos que separan los miles de las centenas y los millones de los miles para evitar confusiones con números decimales; y en segundo lugar, pasar los datos de carácter a numérico, ya que mediante el Web Scraping se guardan los valores como tira de carácteres. Por otro lado, además de la transformación, hemos añadido dos columnas derivadas que nos indican el ratio de población por aeropuerto y el ratio de superficie por aeropuerto respectivamente. También hemos añadido una columna que clasifica el estado según el número de vuelos que ha tenido, lo clasificaremos por muy pequeño, pequeño, mediano, grande y muy grande.

3.3. Comprendiendo nuestros datos

En este apartado ,vamos a entender nuestros datos. Para ello, realizamos una búsqueda filtrando por número de cola de avión para ver cómo se comporta. Como conclusión sacamos que los aviones siguen rutas establecidas entre aeropuertos, información importante para nuestro estudio posterior.

3.4. Análisis exploratorio de los datos

En esta parte para profundizar más en nuestros datos, realizamos un análisis exploratorio donde pudimos ver quiénes fueron los aeropuertos con más y menos vuelos en 2023. En estos podemos destacar el aeropuerto de PUB como el que menos vuelos tiene y el de ATL como el aeropuerto con más vuelos. Por otro lado, representamos quiénes fueron las compañías con el mayor y menor número de vuelos. Podemos destacar la compañía Southwest Airlines como la compañía responsable del mayor número de vuelos y

destacamos también a la compañía GoJet Airlines como la compañía con menor número de vuelos. Luego, estudiamos quiénes son las compañías que trabajan en un mayor número de aeropuertos, siendo SkyWest Airlines la compañía que trabaja en un mayor número de aeropuertos, en la otra cara de la moneda tenemos a la compañía Hawaiian Airlines como la que trabaja en menos aeropuertos. Para finalizar, representamos en un diagrama de tarta las 5 compañías que trabajan en más aeropuertos y, por último, visualizamos las 5 rutas con mayor número de vuelos donde el primer lugar lo ocupa la ruta entre los aeropuertos HNL-OGG.