

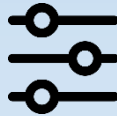





Análisis de Recursos Aeroparque Jorge Newbery



Autor: Isla Nicolás Diego

Aeropuertos Argentina 2000

-  1. Contexto
-  2. Problemática y Objetivos
-  3. Data Acquisition
-  4. Exploración de Datos (EDA)
-  5. Algoritmos y Modelos
-  6. Insights y Recomendaciones

Contexto Comercial

Dada su excelente ubicación geográfica, el Aeropuerto Jorge Newbery es estratégico en la explotación de empresas aéreas tanto domésticas como internacionales. En 2017 fue el aeropuerto con mayor afluencia de pasajeros en Argentina y sus principales rutas de vuelo son domésticas.

Luego de la pandemia Covid19 se decretó que el mismo volvería a ser un aeropuerto internacional, abarcando destinos del Mercosur y países de Sudamérica.

Por tal motivo, se proyecta un incremento exponencial de sus operaciones, aunque aún no se realizaron grandes inversiones en su infraestructura para soportar este incremento. Durante la pandemia se reconstruyó la única pista que este aeródromo posee, pero esa obra se presume que no es suficiente ya que la terminal de pasajeros ha permanecido sin grandes incrementos de su capacidad.

Problemática planteada

Se realizará el estudio de los datos provistos por el explotador aéreo actual a fin de prever los picos de capacidad de todos los subsistemas que integran el aeropuerto y así evitar saturación de los mismos, permitiendo una mejor planificación de los recursos.

Acorde a lo planteado por el operador actual, uno de los principales problemas que genera saturaciones es la demora en el arribo de los vuelos, generando superposiciones. Por este motivo, uno de los focos principales del análisis será intentar **predecir las demoras en los vuelos**.

Se buscará responder a las siguientes preguntas:

- ☐Cuál es la cantidad de vuelos por mes?
- ☐Qué cantidad de pasajeros??
- ☐Cómo se distribuyen los vuelos por día y frecuencia?
- ☐Cómo se aprovechan los recursos de pista?
- ☐Hay vuelos con demoras? Se corresponden con algún origen particular??
- ☐Cuáles son los horario pico de pasajeros en la terminal?



Descripción del Dataset

El Dataset fue provisto por el operador actual del aeropuerto. Cuenta con una gran cantidad de registros (127929) y variables (28), en su mayoría de tipo categóricas.

Se realizó una evaluación de duplicados y valores nulos. Luego de un análisis de los datos y evaluación de su potencialidad, se definió la siguiente estrategia para su tratamiento:

1. Eliminar variables con pocos datos
2. Reemplazo de variables ETA y ATA por STA
3. Reemplazo de Valores Nulos

Se crearon variables nuevas (Delayed, Demora_Min) que mejoran el análisis objetivo, a partir de otras existentes que por si solas no eran tan útiles.

```
RangeIndex: 127929 entries, 0 to 127928
Data columns (total 28 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Aero        127929 non-null object
1   #Vuelo      127929 non-null object
2   CShare      0 non-null      float64
3   Origen      127929 non-null object
4   Via         778 non-null    object
5   STA         127929 non-null object
6   SUG         0 non-null      float64
7   ETA         100446 non-null object
8   ATA         125588 non-null object
9   Tipo        127929 non-null object
10  Asignar     0 non-null      float64
11  Pos         125598 non-null object
12  Ter         127817 non-null object
13  Sec         115477 non-null object
14  Rmk         127929 non-null object
15  Cin         113648 non-null object
16  L&F         189 non-null    float64
17  Pax         125560 non-null float64
18  Vip         98 non-null     object
19  Mat         125738 non-null object
20  Acft        127929 non-null object
21  Obs.        620 non-null    object
22  Aero.1      125598 non-null object
23  #Rot        125598 non-null object
24  Cabecera    125629 non-null float64
25  año         127929 non-null int64
26  mes         127929 non-null int64
27  hora        127929 non-null int64
dtypes: float64(6), int64(3), object(19)
```



Datasets Complementarios

Con el objetivo de enriquecer el análisis, se utilizaron 3 dataset complementarios obtenidos en diferentes repositorios de GitHub:

<https://davidmegginson.github.io/ourairports-data/airports.csv>

<https://davidmegginson.github.io/ourairports-data/countries.csv>

<https://raw.githubusercontent.com/luke/ISO-3166-Countries-with-Regional-Codes/master/all/all.csv>

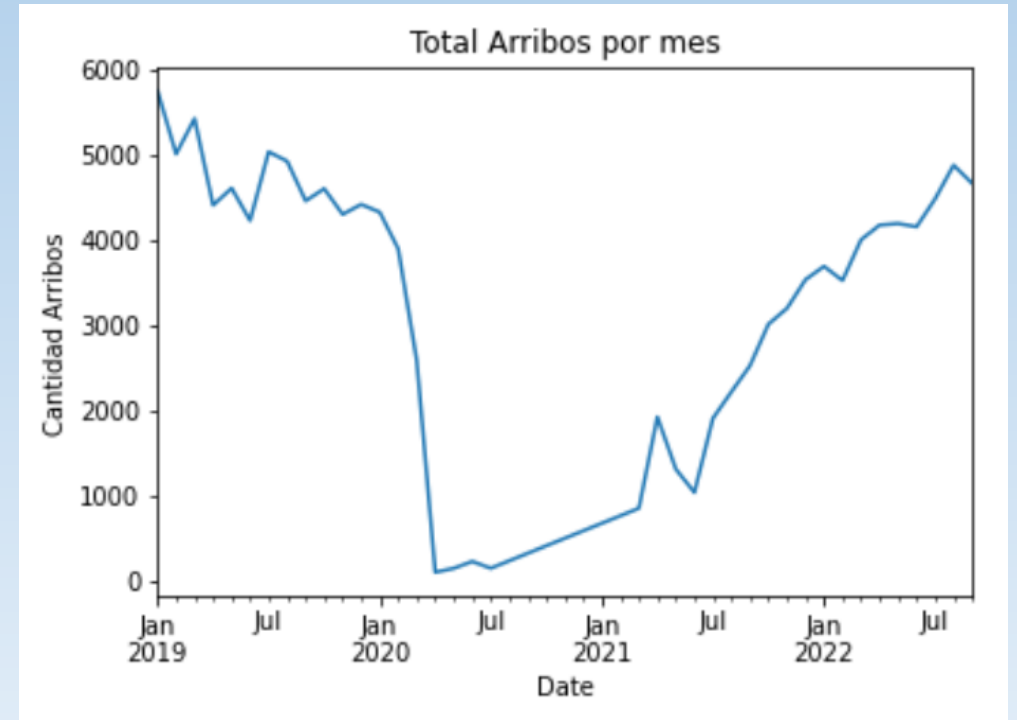
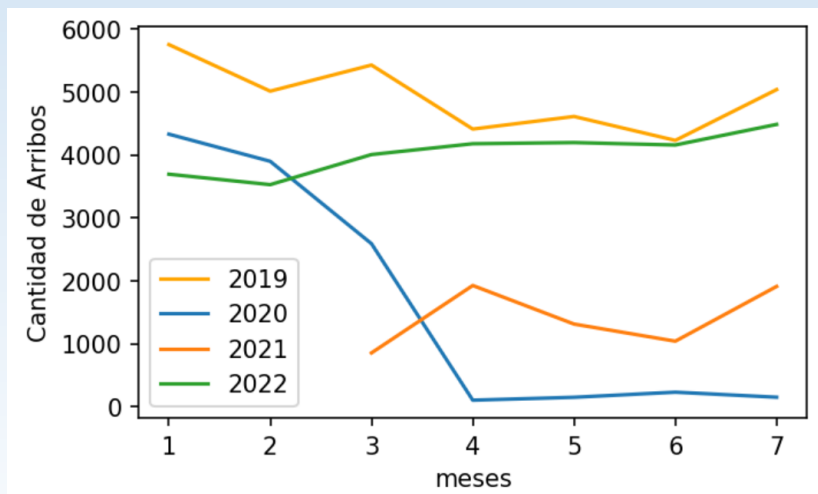
El objetivo principal es enriquecer la información de los aeropuertos de origen de los vuelos para poder analizar su incidencia en las Demoras.

```
Int64Index: 125582 entries, 0 to 125581
Data columns (total 37 columns):
#   Column              Non-Null Count  Dtype
---  -
0   Aero                 125582 non-null object
1   #Vuelo               125582 non-null object
2   Origen               125582 non-null object
3   STA                  125582 non-null datetime64[ns]
4   ETA                  125582 non-null datetime64[ns]
5   ATA                  125582 non-null datetime64[ns]
6   Tipo                 125582 non-null object
7   Pos                  125582 non-null object
8   Ter                  125582 non-null object
9   Sec                  125582 non-null object
10  Rmk                   125582 non-null object
11  Cin                   125582 non-null object
12  Pax                   125582 non-null float64
13  Mat                   125582 non-null object
14  Acft                  125582 non-null object
15  Aero.1                125582 non-null object
16  #Rot                  125582 non-null object
17  Cabecera              125582 non-null float64
18  año                    125582 non-null int64
19  mes                    125582 non-null int64
20  hora                   125582 non-null int64
21  HOUR                   125582 non-null int64
22  WEEKDAY                125582 non-null int64
23  Convert                125582 non-null timedelta64[ns]
24  Demora_Min             125582 non-null float64
25  Delayed                 125582 non-null int64
26  type                    123826 non-null object
27  name                    123826 non-null object
28  municipality           123825 non-null object
29  country_x               123826 non-null object
30  iso_country             123826 non-null object
31  region                  123826 non-null object
32  sub-region              123826 non-null object
33  latitude_deg            123826 non-null float64
34  longitude_deg           123826 non-null float64
35  wikipedia_link_x       112881 non-null object
36  local_code              114354 non-null object
dtypes: datetime64[ns](3), float64(5), int64(6), object(22), timedelta64[ns](1)
```

Cantidad de Vuelos por mes

Se puede observar la marcada reducción en la cantidad de vuelos durante el año 2020 producto de la pandemia y cómo se fue recuperando fuertemente la actividad a partir de julio de 2021

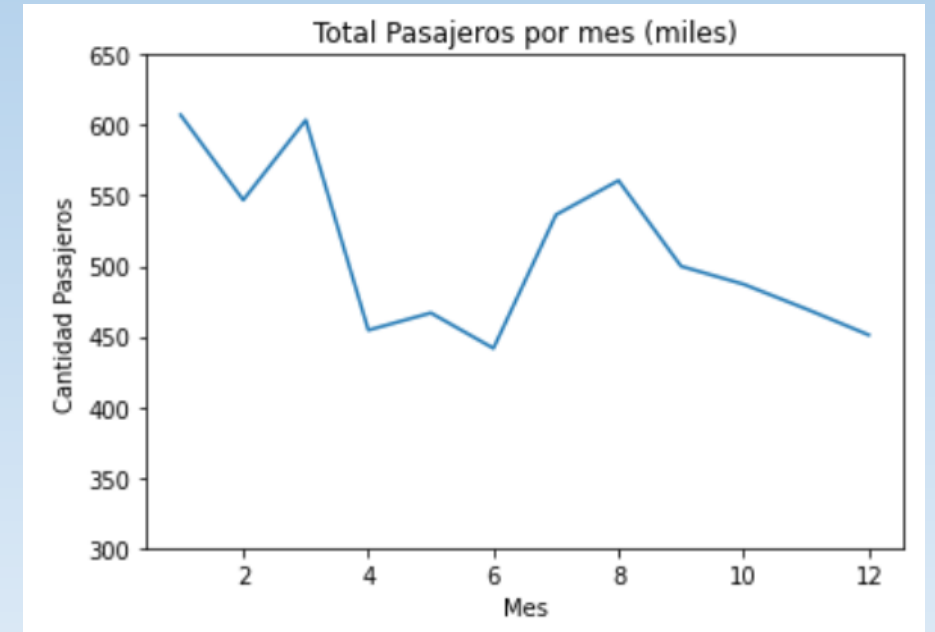
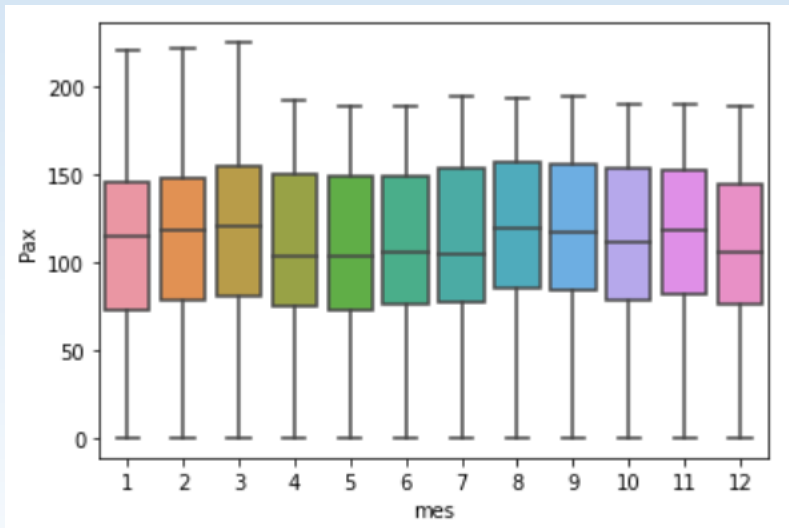
Para poder comparar los niveles de ocupación de recursos actuales, se decide comparar los períodos de Enero a Julio ya que son los últimos datos disponibles para 2022



Podemos observar que los niveles de operación en cuanto a cantidad de vuelos para 2022 aún no superan los niveles de 2019, lo cual **permite suponer que aún no se saturó** la capacidad operativa del aeropuerto.

Cantidad de Pasajeros por mes

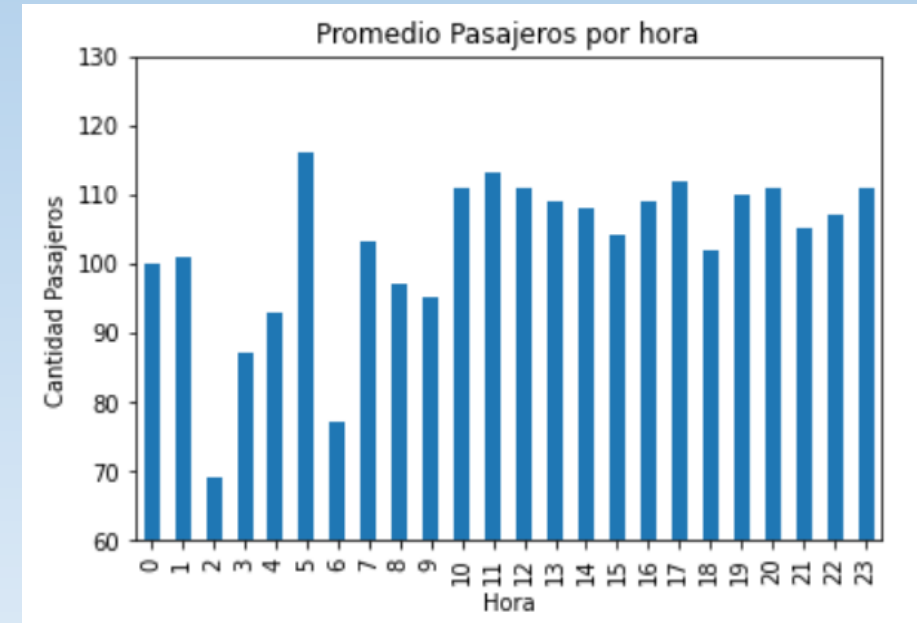
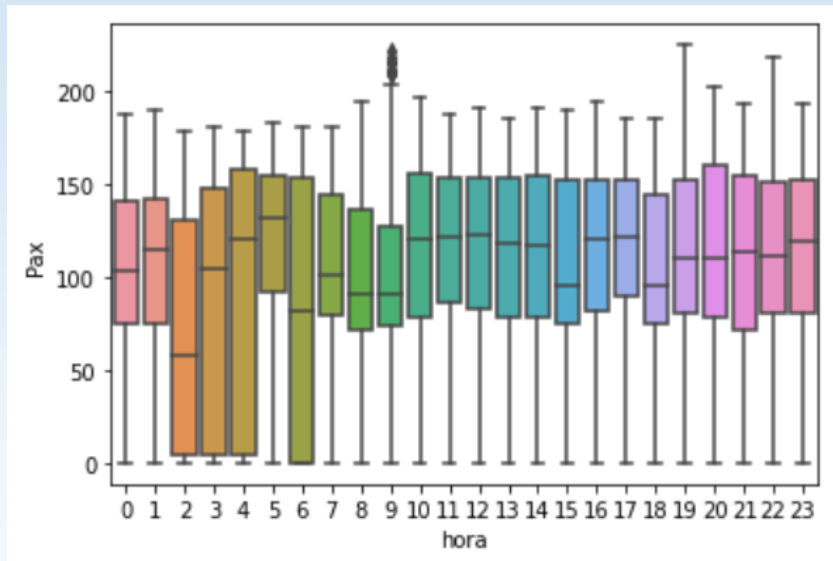
Para analizar la estacionalidad mensual de pasajeros se decide tomar como referencia el año 2019 que es el más completo y representativo (no está afectado por la pandemia). Se observan picos de capacidad en los meses de vacaciones, tanto de verano como de invierno.



Si analizamos la cantidad de pasajeros por vuelo en los diferentes meses del año podemos observar que tienen distribuciones similares, lo que no sugiere que haya estacionalidad en cuanto a la ocupación de los vuelos (vuelos significativamente más llenos o vacíos según la época del año)

Cantidad de Pasajeros por hora

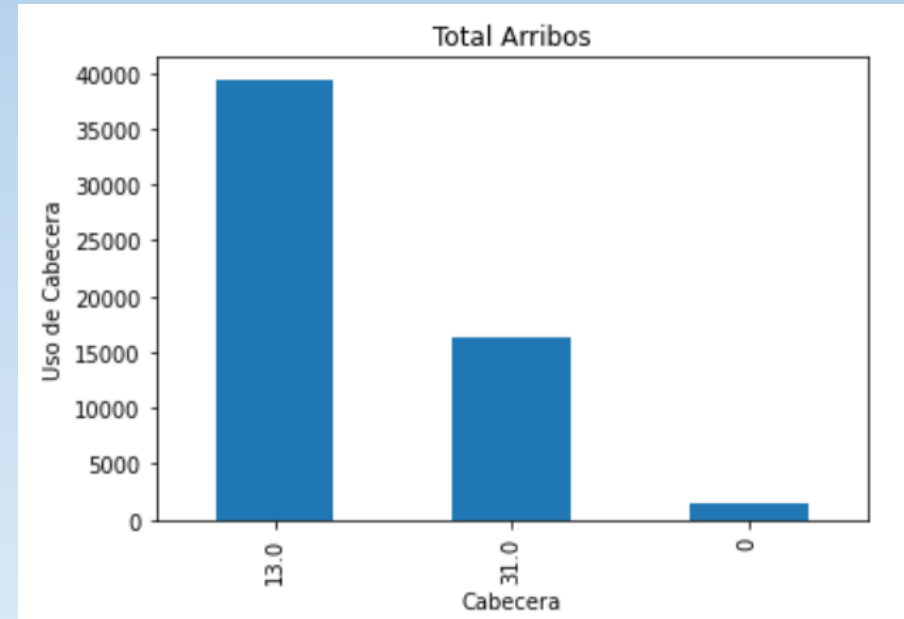
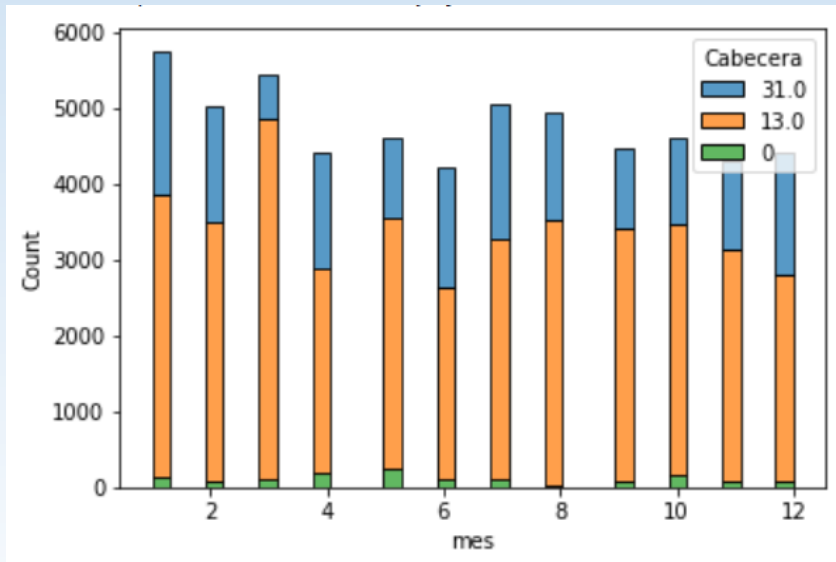
En cuanto a la distribución horaria de pasajeros promedio por día, podemos observar que en los horarios de madrugada y mañana los vuelos suelen traer menos pasajeros, lo cual permite suponer una cierta capacidad ociosa en el aeropuerto, aunque esto dependerá de los recursos asignados en dichas franjas horarias (en general suele haber menos personal).



También podemos observar que en los horarios de madrugada la dispersión de la cantidad de pasajeros por vuelo aumenta, teniendo más variabilidad y vuelos con muy pocos pasajeros.

Uso de Cabecera

Si bien el aeropuerto cuenta con sólo una pista de aterrizaje, la misma puede ser utilizada en 2 sentidos, lo que determina que hayan 2 “cabeceras” posibles. Para los vuelos del año 2019 se observa que una fue significativamente más utilizada que la otra.

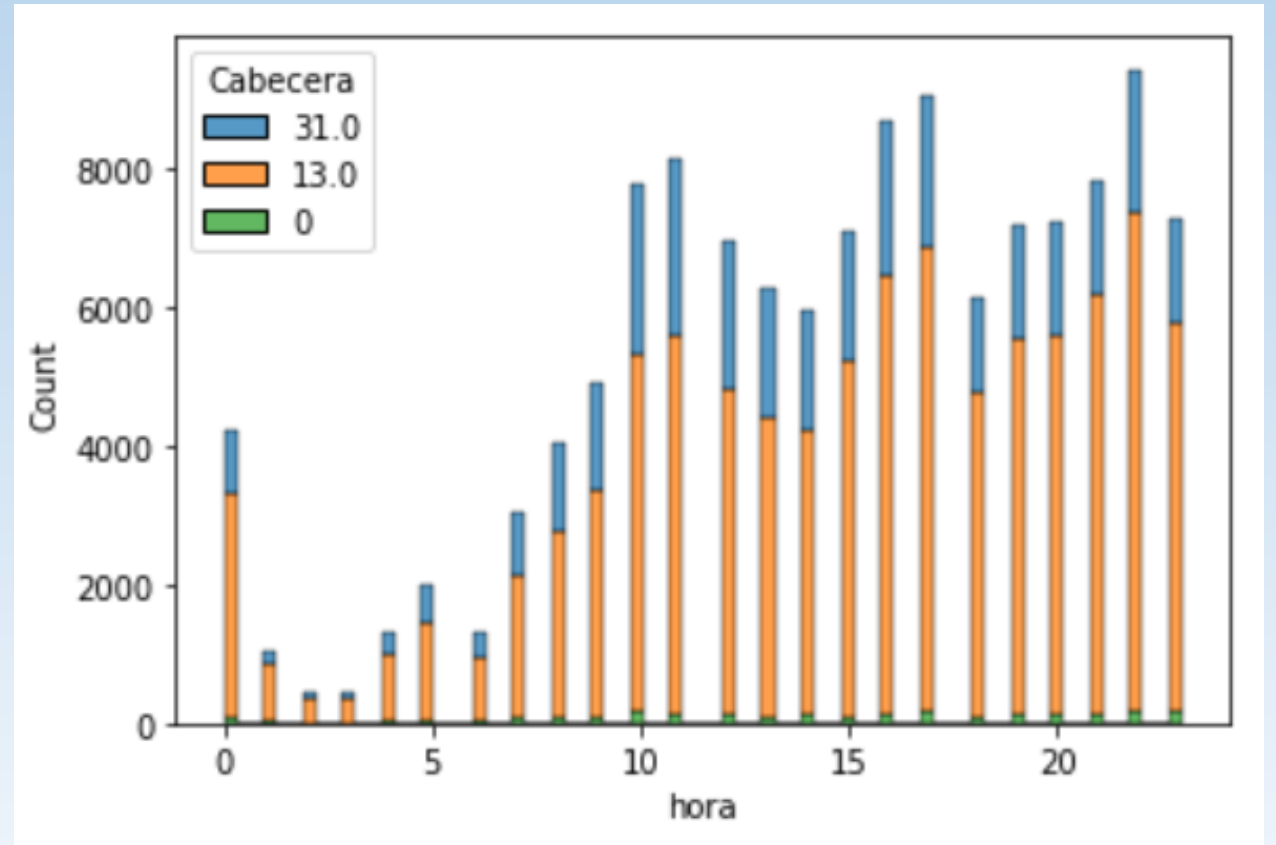


Al observar la distribución de vuelos mensual observamos que la proporción entre usos de cabecera es similar todos los meses. Se desconoce si existe algún impedimento o restricción, pero a priori podría representar una oportunidad para optimizar recursos.

Uso de Cabecera

Si observamos la distribución horaria de los vuelos de todo el período (2019-2022) podemos ver que la proporción de uso de cabeceras es similar en los diferentes rangos, siempre menor en cabecera 31.

Por otra parte, podemos ver que la cantidad de vuelos recibidos en horarios de madrugada es considerablemente menor que el resto del día, lo que supone una gran capacidad operativa ociosa/potencial.

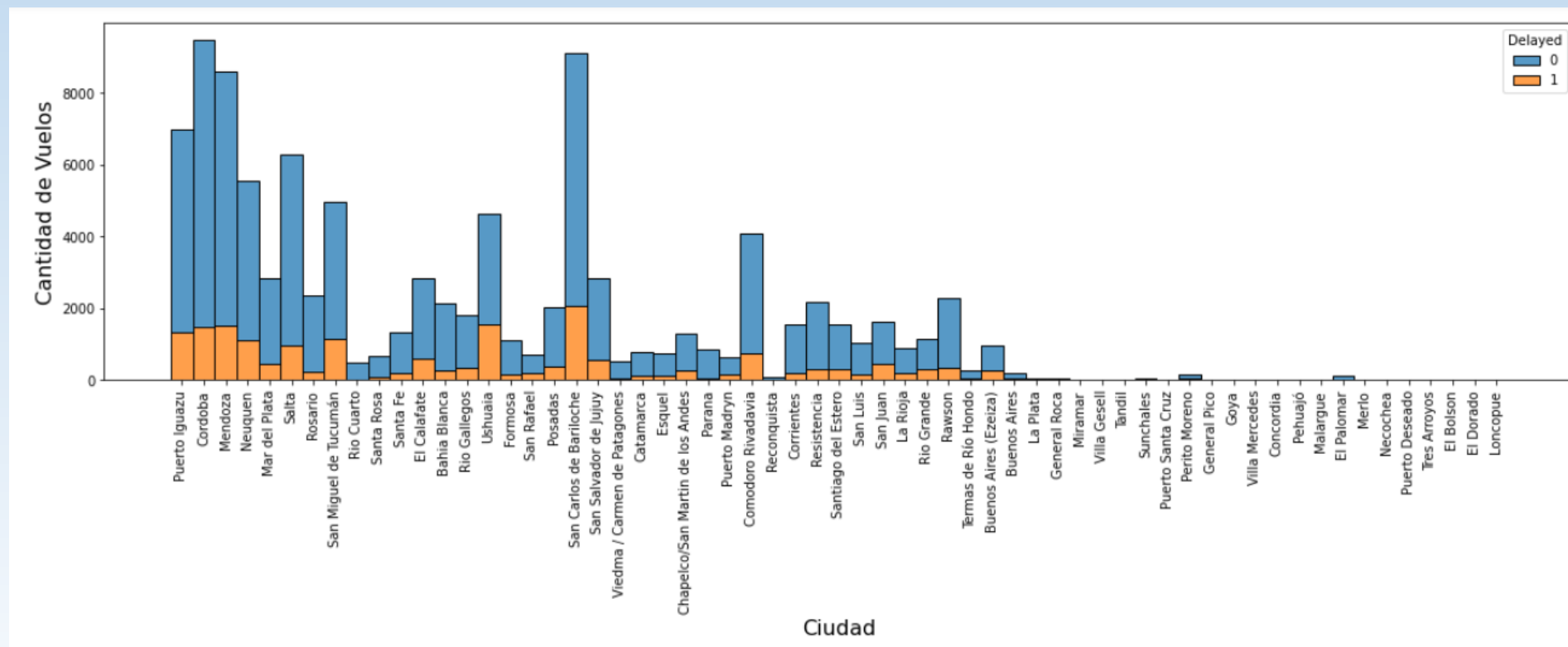


Demoras por Origen

El **87%** de los vuelos recibidos provienen de **Argentina**, de los cuales el **25%** presenta **demoras** en su horario de arribo.

Aeropuertos con mayor cantidad de vuelos demorados:

- Bariloche
- Córdoba
- Mendoza
- Ushuaia
- Puerto Iguazú





Definición de Modelos

Se decidió desarrollar dos modelos para trabajar sobre la problemática de las demoras en los vuelos y su impacto en la asignación de los recursos aeroportuarios.

El primer modelo es de **clasificación** y su objetivo es intentar predecir si un vuelo llegará a destino “En horario” o “Demorado”. En base a esas dos categorías se aplicaron tres algoritmos y se evaluó su desempeño.

El segundo modelo es de **regresión** y su objetivo es intentar predecir los minutos de desvío en la llegada de cada vuelo, es decir, los minutos que llegará demorado o antes del horario de arribo previsto (STA). Se aplicaron



Modelo I

El primer modelo es de **clasificación** y su objetivo es intentar predecir si un vuelo llegará a destino “En horario” (0) o “Demorado” (1).

En base a esas dos categorías se aplicaron tres algoritmos y se evaluó su desempeño.

| Algoritmo | Variable | Accuracy | Precision | Recall | F1 Score |
|--------------------------|----------|----------|-----------|--------|----------|
| KNN | 0 | 0,80 | 0,82 | 0,97 | 0,89 |
| | 1 | | 0,37 | 0,07 | 0,11 |
| Random Forest | 0 | 0,81 | 0,81 | 1,00 | 0,90 |
| | 1 | | 0,59 | 0,02 | 0,03 |
| Random Forest Balanceado | 0 | 0,66 | 0,86 | 0,68 | 0,76 |
| | 1 | | 0,28 | 0,53 | 0,37 |
| Regresión Logística | 0 | 0,81 | 0,81 | 1,00 | 0,90 |
| | 1 | | 0,00 | 0,00 | 0,00 |

El modelo presenta un gran desbalance de datos, lo que se traduce en **altos valores de Accuracy** pero **baja performance en F1 Score**. Esto se debe a que en general predice bien la variable “En horario”= 0 que es la que se encuentra más presente, pero no logra buen desempeño en la variable “Demorado” = 1.

Para la variable “Demorado” es deseable minimizar los Falsos Negativos, ya que implicarían predecir erróneamente un vuelo en tiempo. Por esto debemos prestar especial atención a los valores de **Recall**.

La performance de los 4 algoritmos es baja, pero podríamos decir que el Random Forest Balanceado es el que mejores métricas tiene en la predicción de vuelos Demorados vs En horario.



Modelo II

El segundo modelo es de **regresión** y su objetivo es intentar predecir los minutos de desvío en la llegada de cada vuelo, es decir, los minutos que llegará demorado o antes del horario de arribo previsto (STA).

| Algoritmo | RMSE | RAE | R2 |
|------------------|-------|------|------|
| Random Forest V1 | 11,44 | 0,89 | 0,12 |
| Random Forest V2 | 7,74 | 0,63 | 0,60 |
| XGBoost | 7,62 | 0,62 | 0,61 |

Al utilizar el algoritmo Random Forest (Versión 1) obtenemos un RAE alto (cercano a 1) y un muy bajo R2 (cercano a 0), lo que nos muestra que **el algoritmo tiene una performance demasiado baja** para poder ser considerado.

Si volvemos a correr el algoritmo agregando la variable “Delayed” (Versión 2), observamos una notable **mejora en el R2** aunque los **valores de RAE siguen siendo altos** como para poder estar satisfechos con el modelo.

Por último, utilizamos el algoritmo XGBoost para las mismas variables que el Random Forest V2 y logramos obtener una **mejora en el RAE, sosteniendo los valores de R2**.



Cantidad de vuelos

En el año 2022 se observa una tendencia creciente de vuelos que plantea la necesidad de ampliar recursos, aunque en el corto plazo no sería un problema dado que **no se han alcanzado aún los niveles de operación del año 2019**. Una posible opción sería analizar la posibilidad de **balancear la cantidad de vuelos** que son recibidos en cada cabecera, para mejorar la utilización del recurso (no saturar uno en favor del otro).

Por otra parte, se observa una **gran capacidad ociosa en los turnos de madrugada** respecto al resto del día. Esto podría aprovecharse **redireccionando vuelos desde otros aeropuertos**, quizá de nuevos destinos en los cuales dicha franja horaria sea más beneficiosa para su combinación. De esta forma se podría aprovechar ese recurso y, en tal caso, direccionar vuelos de la franja horaria central/tarde hacia el otro aeropuerto presente en la ciudad.



Cantidad de pasajeros

Se observa que en los horarios de madrugada, así como hay menos vuelos también suelen tener menor cantidad de pasajeros. Sería recomendable **reasignar rutas de vuelo en esos horarios y tratar de acompañarlo con un plan de incentivos/beneficios para los pasajeros que decidan volar en esa franja**; el aeropuerto debería ofrecer beneficios a las empresas aéreas para que pudieran transmitirse a los usuarios finales.

Predicción de Demoras

Según lo analizado, podemos concluir que la información provista por el operador **no cuenta con las variables relevantes suficientes para generar el modelo de predicción buscado**. Las variables presentes no tienen la incidencia suficiente en la demora de un vuelo, lo que dificulta generar modelos precisos en sus predicciones.

Para continuar investigando se recomienda intentar conseguir otras variables que puedan tener una correlación más directa, como por ejemplo la distancia recorrida por el vuelo, condiciones meteorológicas (al despegar, durante el vuelo, al aterrizar), experiencia de los pilotos, etc.