



## Facultad de Ingeniería

9558 - CIENCIA DE DATOS

Primer Cuatrimestre de 2024

ANÁLISIS EXPLORATORIO

Marcela Jazmin Cruz  
110066

Gualdi Martina  
110513

<b>Introducción.....</b>	<b>2</b>
<b>Estudio a lo largo del tiempo.....</b>	<b>3</b>
A lo largo de los meses.....	3
<b>Estudio de via de compra de vuelos.....</b>	<b>9</b>
¿Cuál es el medio de compra más común?.....	9
¿Tienen alguna conexión los días de anticipación?.....	10
¿Tiene alguna conexión con la cantidad de personas que viajan?.....	11
<b>Estudio sobre los pasajeros.....</b>	<b>13</b>
<b>Estudio tipo de vuelos y servicios.....</b>	<b>14</b>
Respecto a viajes individuales.....	15
<b>Visualización particular.....</b>	<b>16</b>

## Introducción

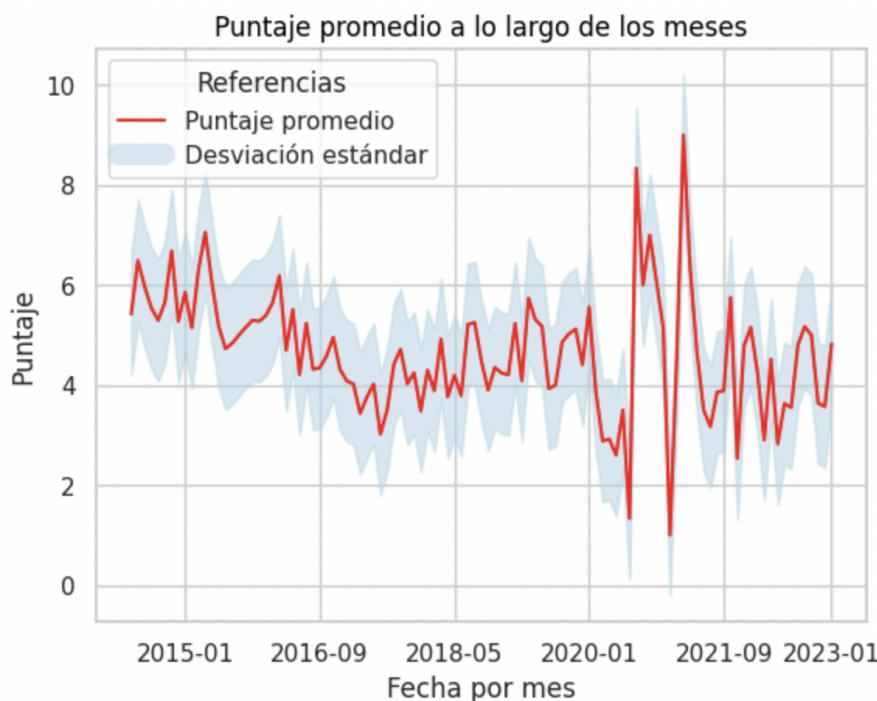
En análisis exploratorio realizado en el presente trabajo se llevó a cabo en base a los siguientes interrogantes:

- Las personas que viajan solas seguro son de seleccionar extras como:
  - wants\_preferred\_seat
  - wants\_in\_flight\_meals
  - wants\_extra\_baggage
- ¿Eligen un solo extra o todos? ¿Prefieren alguno en particular?
- Las personas que viajan solas compran el vuelo o planean su viaje con más anticipación que las que viajan en grupo?
- A mayor duración de horas vuelo, ¿es más tarde la hora de salida? ¿Se busca viajar de noche para aprovechar las horas de sueño?
- La persona que viaja sola tiende a comprar vía mobile y si viaja en grupo vía internet?
- ¿Cuál es la vía más común?
- La pandemia tuvo alguna consecuencia en los vuelos?
- ¿Cómo fueron las estadísticas a lo largo del tiempo? (meses y años)

# Estudio a lo largo del tiempo

## A lo largo de los meses

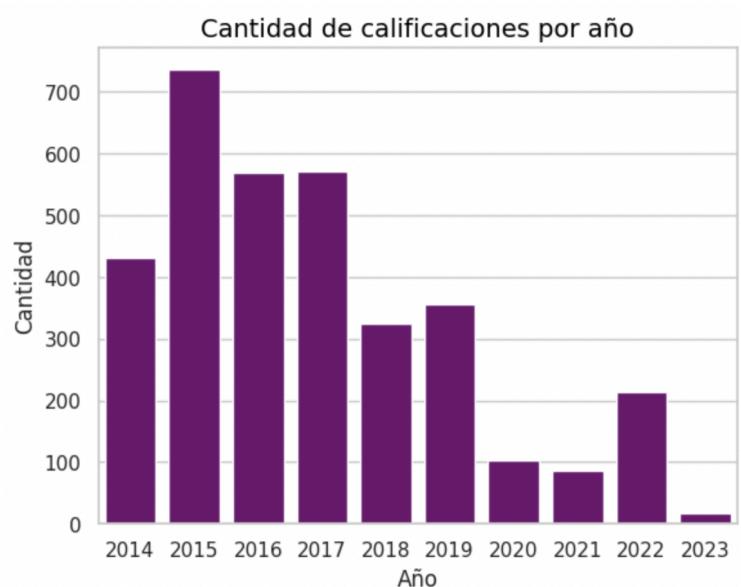
Uno de los objetivos del presente trabajo fue analizar cómo cambiaba la opinión de los pasajeros a lo largo del tiempo, por lo cual se decidió hacer el siguiente gráfico, el cual evalúa las calificaciones promedio mes a mes desde el 05/2014 hasta el 01/2023:



Como bien podemos observar se notan picos en distintos períodos de tiempo como por ejemplo entre el 2020 y el 2021, como también, en el mismo periodo, se puede observar que la desviación estándar del promedio es mucho menor a otros momentos. Esto puede deberse a distintas cosas, principalmente: hubo pocas evaluaciones por lo tanto el análisis es menos preciso o el servicio realmente fue positivo por lo cual todas las calificaciones rondan los mismos valores.

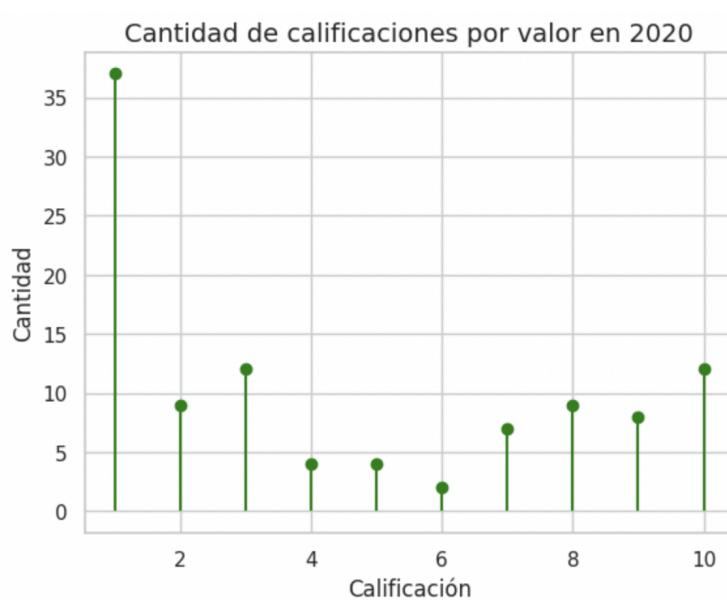
Lo mencionado anteriormente nos llevó a realizar un estudio más profundo sobre el impacto de la pandemia en los valores de las aerolíneas.

En primer lugar, nos parece primordial analizar cómo va cambiando la cantidad de comentarios que se realizan año a año. Mientras que en 2015 (año en el cual tenemos la mayor desviación estándar por lo visto en el gráfico



anterior) se realizaron más de 700 calificaciones, en el año 2020 la cantidad de comentarios cayó abruptamente con un valor de 100 devoluciones por parte de los pasajeros. Estos valores confirman una de nuestras teorías de porqué ese periodo tiene mucha menor desviación estándar.

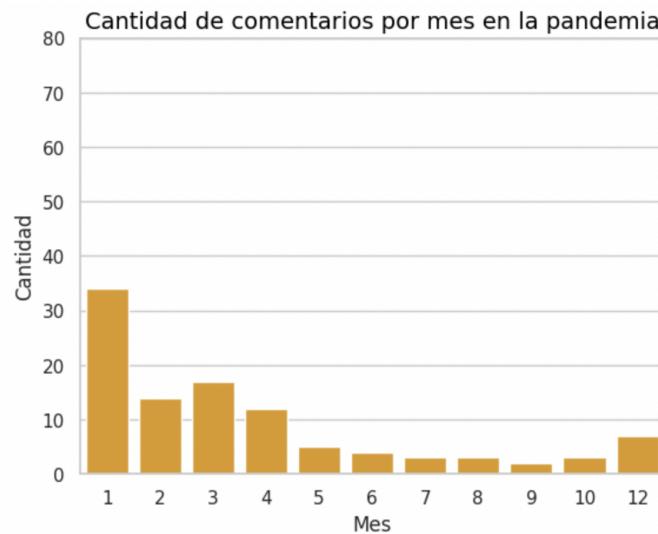
Continuando con la segunda de nuestras hipótesis (buen servicio => la mayoría son buenas calificaciones), con el siguiente gráfico podemos mostrar que la misma no es del todo correcta. El gráfico muestra cuántas calificaciones por valor hubieron en el año pandémico. Si bien la cantidad de comentarios no son muchos (no superan los 100) podemos notar una amplia diferencia entre los que puntuaron 1 y el resto de las valoraciones. Por lo visto en el gráfico hay más de 35 valoración en 1 y menos de 15 en valoración 10, los valores intermedios fluctúan entre 0 y 14 repeticiones (no alcanzando ninguno de los dos extremos).



De esta manera podemos decir que la segunda hipótesis de porque la desviación estándar en el periodo del 2020 es mas chica, es incorrecta ya que en términos de cantidades hay muchos mas comentarios que fueron evaluados como negativos que los que fueron evaluados como excelentes, sin embargo la poca desviación puede deberse a que las puntuaciones entre 2 y 10 toman valores bastante similares, o mejor dicho, no tienen cambios muy abruptos.

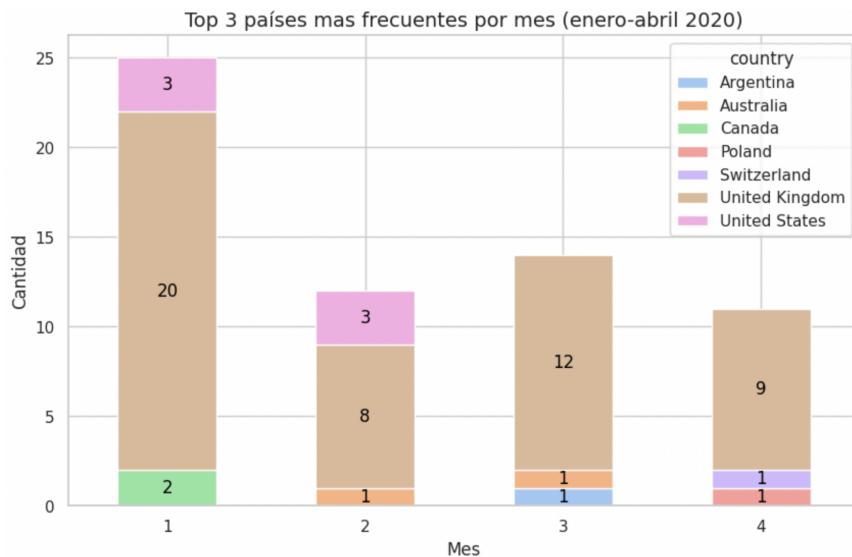
Como bien pudimos observar en la primera hipótesis consideramos interesante profundizar en el análisis de qué fue lo que pasó a lo largo de los meses del 2020. El gráfico muestra la cantidad de comentarios que hubieron en los 12 meses de ese año.

Desde el mismo podemos sacar distintas conclusiones:



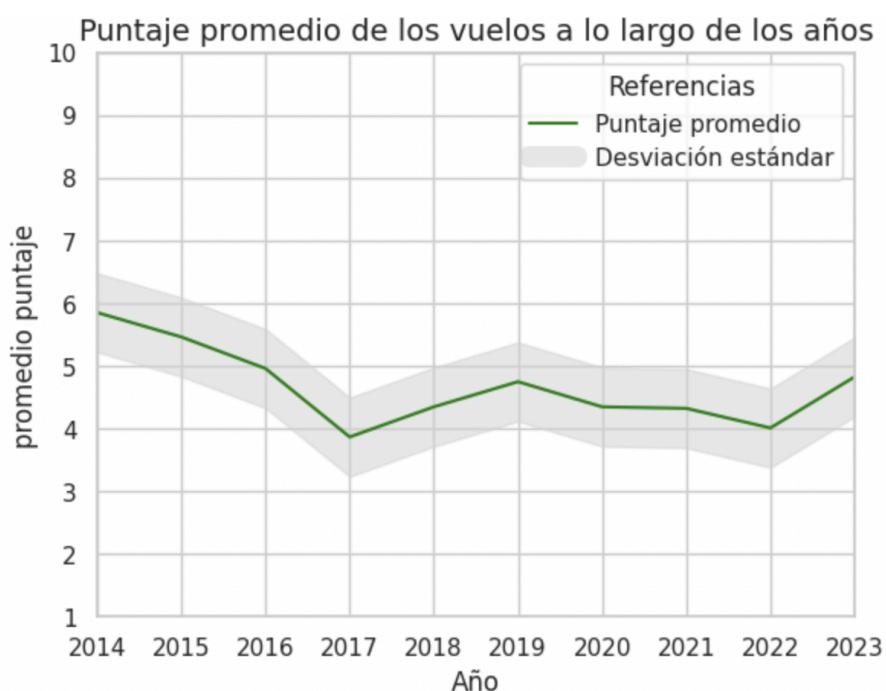
1. El mes 1 (Enero) puede tener la mayor cantidad de comentarios ya que aun no estaba “desatada” del todo la pandemia por lo cual la frecuencia de vuelos era la usual.
2. A medida que avanzan los meses, al igual que la catástrofe, podemos tener en consideración que los meses 2, 3 y 4 (febrero, marzo y abril) fueron los meses de repatriación de aquellas personas que habían quedado varadas en distintas partes del mundo a causa del cierre de fronteras. Puede deberse a este motivo que la cantidad de comentarios haya bajado (porque la cantidad de vuelos así lo hizo) pero que no haya sido un cambio muy brusco.
3. Del mes 5 al 9 (de mayo a septiembre) fueron los momentos más críticos en muchas partes del mundo, donde todo estaba cerrado, por lo cual tiene mucha lógica que la cantidad de comentarios disminuya notablemente.
4. Finalmente hacia fin de año (sacando de consideración el mes de noviembre del cual no se nos brinda información sobre ese año) podemos ver como vuelve a subir la tendencia de cantidad de comentarios, lo que puede deberse a varias cosas: apertura de fronteras y más viajes, viajes por causas festivas en el mes de diciembre, pérdida de miedo a viajar luego de tantos meses de pandemia crítica.

Hemos decidido profundizar las conclusiones mencionadas anteriormente para analizar si puede ser considerado así o no. En el gráfico “Top 3 países más frecuentes por mes (enero-abril 2020)” podemos observar, como bien dice su título, cuáles fueron los países de origen más frecuentes en ese momento. Si bien nuestra primera conclusión fue considerar que en el mes 1 los viajes aún se realizaban con normalidad, podemos comenzar a notar una tendencia hacia la frecuencia de un país en particular. Así mismo, los siguientes tres meses no fueron muy cambiantes. La hipótesis con respecto a los mismos era que los países más frecuentes eran los países con mayor crisis por la pandemia ya que relacionamos esa crisis con los vuelos de repatriación. Más allá de que muchos de los países más críticos, como lo fueron Italia y España, no aparecen en el gráfico, el denominador común de los cuatro meses es que el más frecuente fue United Kingdom el cual también tuvo un momento muy crítico en esa época del 2020 y puede haber tenido muchos vuelos de repatriación para esa altura.



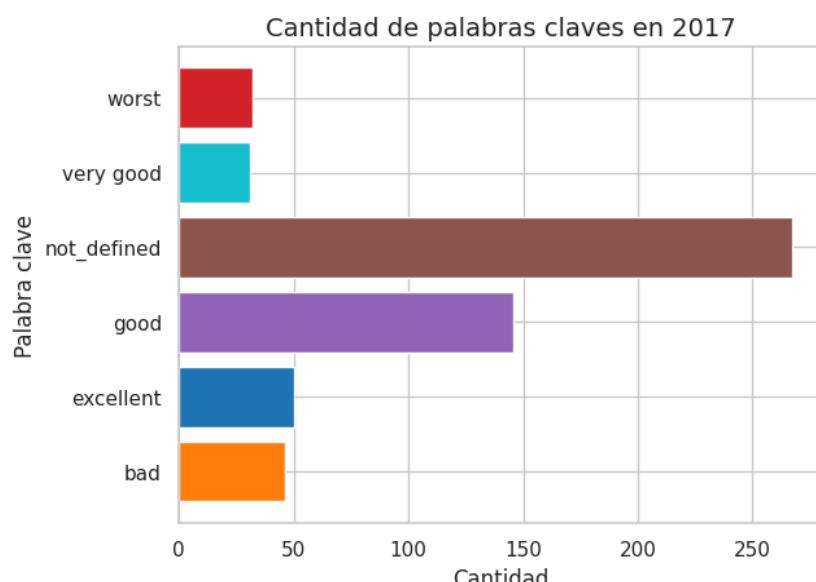
## A lo largo de los años

Dejando de lado la pandemia, nos pareció oportuno para la investigación analizar el flujo de los años al igual que lo hicimos con los meses. En el siguiente gráfico podemos observar el puntaje promedio a lo largo de los años y su desviación estándar. En este caso la desviación no desprende mucho análisis ya que es bastante parejo en todo el recorrido. Sin embargo, podemos hacer foco en otros dos puntos, el 2014 (promedio más alto de todos) y 2017 (promedio más bajo de todos).



En el año 2017 podemos observar que la cantidad de comentarios a lo largo de los meses es bastante pareja y alta, sin embargo eso no nos da indicio de porque fue el año con menor calificación, por ese motivo decidimos hacer foco en cuantos comentarios negativos y positivos hubieron en ese año.

En la gráfica “Cantidad de palabras clave en 2017” podemos ver que aunque el número de comentarios con palabra clave “bad” y “worst” es considerable, la gran mayoría de los comentarios no tienen una palabra clave definida y muchos otros tienen como palabra clave “good”. De esta manera no habría muchos indicios de porque el 2017 fue el año con peor promedio de clasificaciones. Consecuentemente, estos resultados nos llevaron a pensar que capaz los comentarios malos tenían un puntaje muy bajo por lo que el promedio se vio vulnerado y no termina siendo tan representativo.



Ahora en el gráfico “puntaje por palabra clave en 2017” podemos apreciar lo que suponíamos anteriormente:



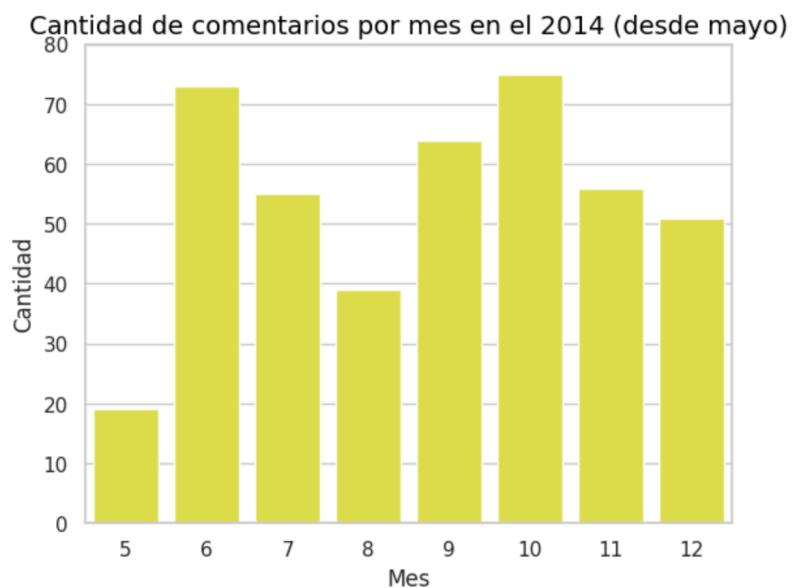
1. Los puntajes de “bad” y “worst” son muy bajos, mayormente rondan entre los valores de 1 y 3.

2. Hay algunas valoraciones con palabra clave “excellent” que igualmente tienen puntaje muy bajo, de esta manera aunque su palabra sea positiva su promedio va a tender a ser bajo.

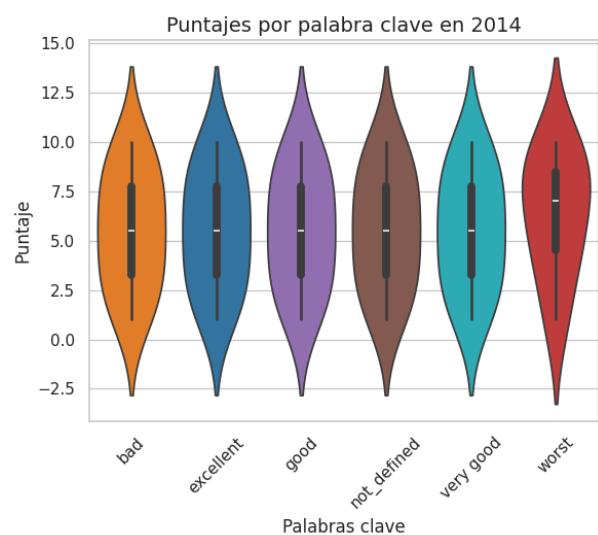
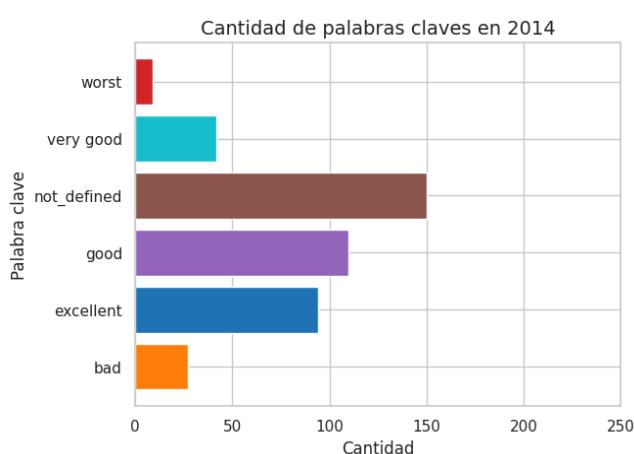
***Estos dos puntos pueden generar que el puntaje promedio anual sea bajo aunque muchos comentarios tengan palabras clave positivas.***

De la misma manera buscamos analizar el caso del 2014, que por lo visto en el gráfico de línea anual era el año con mayor puntaje promedio.

En este caso tenemos la cantidad de comentarios por mes, nótese que comienza desde mayo lo cual podría generar un cálculo menos preciso que otros años en los cuales se tomaron en consideración los 12 meses del año. Podemos notar que el mes de mayo fue el que menos comentarios tuvo, lo que puede deberse a la fecha de inicio de registro de los comentarios de los pasajeros.



Para analizar si el promedio de puntaje es adecuado a los valores de dicho año hicimos el mismo análisis que en el caso anterior. En este caso podemos observar que, aunque muchos de los comentarios son indefinidos, las palabras “good”, “very good” y “excellent” tienen muchos más comentarios que las palabras cuya connotación es negativa, por lo tanto podría ser un indicio de que el promedio es correspondiente.



Así mismo, en el gráfico “Puntajes por palabra clave en 2014” podemos notar que en el caso de los comentarios negativos (en el caso de la palabra “worst” por ejemplo) son más los que tienen puntaje alto que los que tienen puntaje bajo, de esta manera podemos deducir que capaz las palabras claves no son tan consecuentes con el principal foco del comentario de los pasajeros.

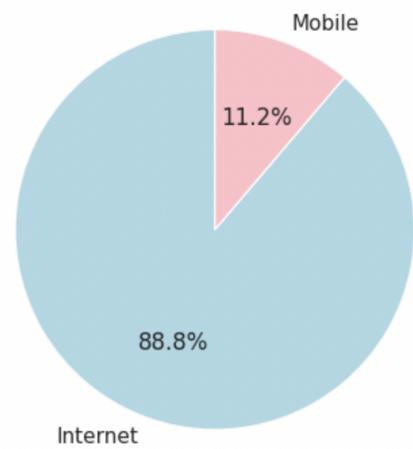
## Estudio de via de compra de vuelos

### ¿Cuál es el medio de compra más común?

A medida que avanzamos en el análisis del dataset, nos pareció interesante ver cómo las personas solían comprar sus vuelos.

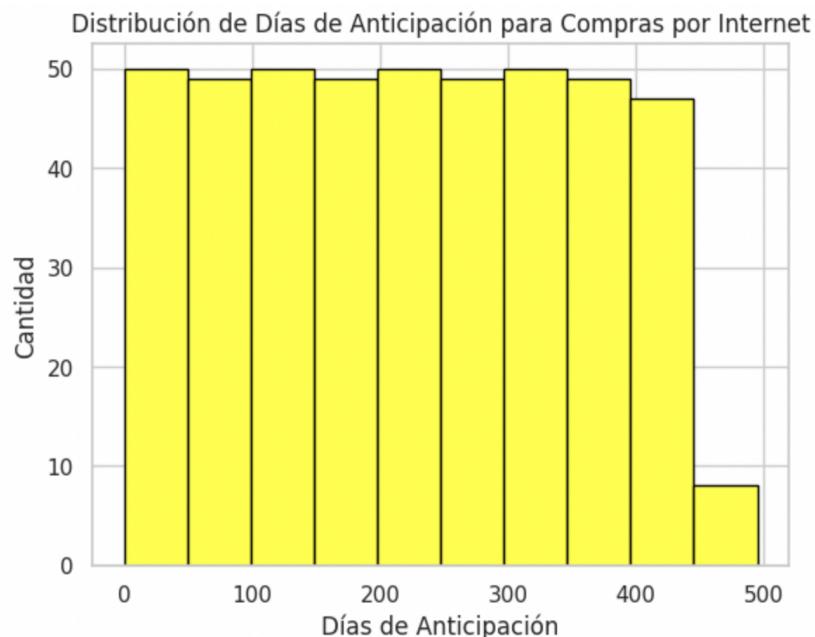
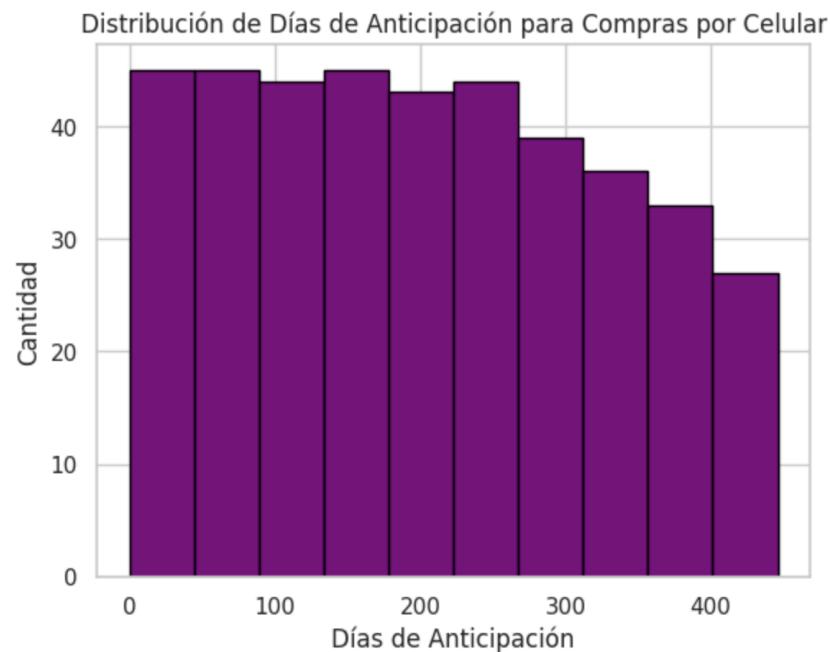
Para comenzar decidimos verificar una de nuestras hipótesis, “Las personas suelen comprar más vuelos por internet que vía celular”. En este gráfico podemos apreciar una notable diferencia entre la cantidad de vuelos que fueron comprados por internet y por teléfono. Esto puede deberse a distintos factores, pero consideramos que uno de los principales puede ser la confianza, ya que suele pasar que a las personas les da más seguridad hacer las cosas por sí mismos y ver personalmente los datos que hacerlo vía telefónica.

Distribucion de vuelos comprados via Internet y via Mobile



## ¿Tienen alguna conexión los días de anticipación?

Otra de nuestras hipótesis se basaba en la idea de que las personas que compraban vía telefónica capaz lo hacen más sobre el momento que las que lo hacen vía internet.

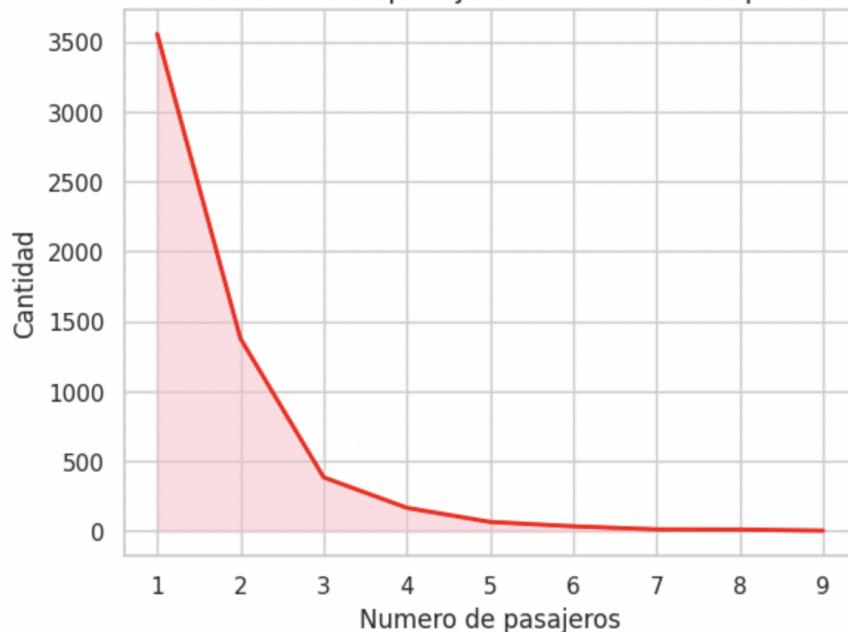


En estos gráficos podemos notar una diferencia lo cual podría confirmar nuestra teoría. Mientras que en las compras de celular la cantidad de vuelos que se compran va disminuyendo a medida que aumentan los días de adelanto, en el caso de internet eso no sucede. En el segundo caso, los valores son bastante constantes (salvo por los días de anticipación mayores a 400 ya que de por sí un viaje no puede reservarse con más de 11 meses de anticipación al regreso).

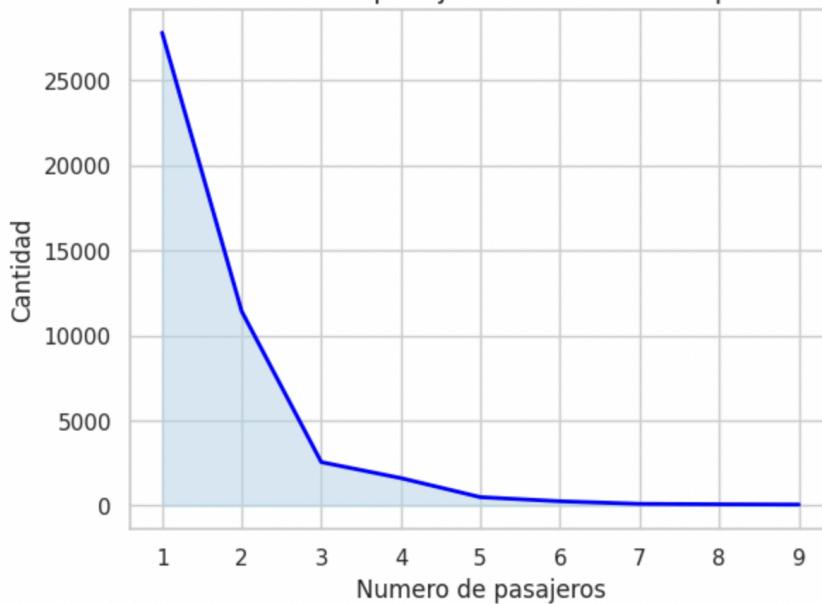
## ¿Tiene alguna conexión con la cantidad de personas que viajan?

Como una de nuestras interrogantes hacía referencia a que si compraban vía mobile era porque los vuelos eran de menos personas, nos pareció interesante hacer un análisis más específico sobre eso.

Distribucion de cantidad de pasajeros en vuelos comprados via Mobile



Distribucion de cantidad de pasajeros en vuelos comprados via Internet



En los gráficos podemos ver la cantidad de vuelos comprados para X cantidad de personas para cada una de las vías de compra. Si bien ambos parecen muy similares no lo son. Mientras que el de vía internet alcanza valores superando los 25.000 los de vía Mobile no superan los 4.000.

Nuestro interrogante podría dividirse en dos partes:

1. Las personas que viajan solas suelen comprar vía “mobile”.
2. La mayoría de los viajes por celular suelen ser de viajes individuales.

Si analizamos cuántos viajes individuales para cada una de las vías hay lógicamente hay muchos más vía Internet. Pero, en términos de porcentaje frente a su total ¿cuál es mayor? Teniendo en cuenta los totales del primer gráfico de la presente sección podemos hacer ese cálculo:

- **Vía Internet:** Si del total de 44366 vuelos comprados vía internet, 27790 representa un 62% del total.
- **Vía Mobile:** Si del total de 5616 vuelos comprados vía internet, 3556 representa un 63% del total.

De esta manera podemos comprobar que no hay una amplia diferencia entre los porcentajes de uno y de otro, por lo que la primera parte nuestra hipótesis no queda comprobada.

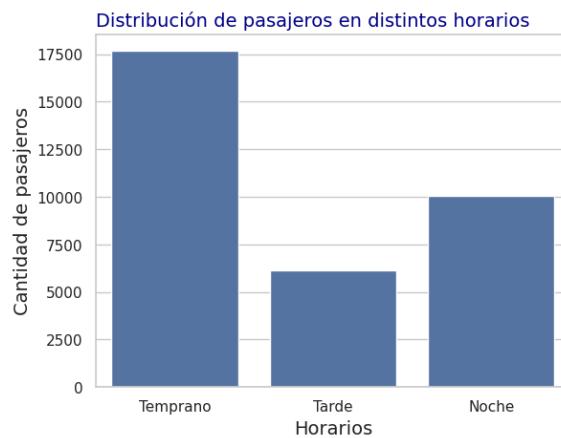
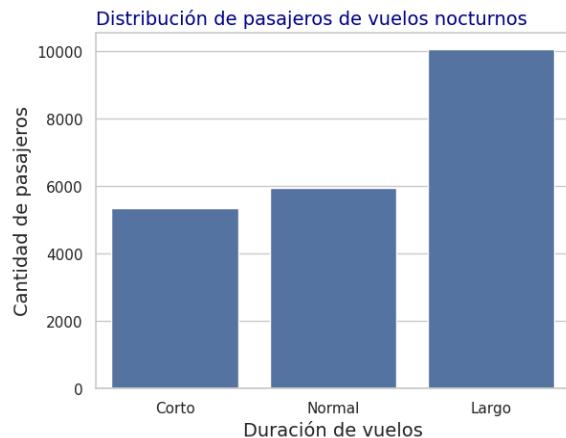
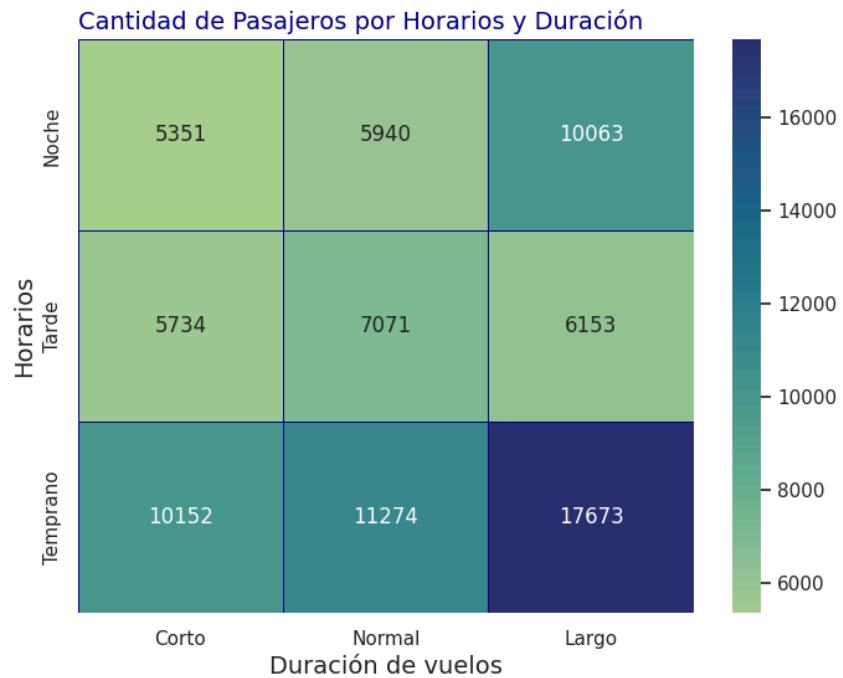
Siguiendo con la segunda parte de la hipótesis, con el gráfico anterior podemos ver que si se cumple que la mayoría de los viajes comprados vía celular son viajes individuales (de hecho el 63% mencionado anteriormente), sin embargo en el caso de vía Internet sucede lo mismo, por lo que no lo podemos comprobar que la vía de compra de los vuelos y la cantidad de pasajeros tengan alguna conexión significativa.

 análisis.ipynb

## Estudio sobre los pasajeros

Con la idea de observar si los pasajeros tenían una tendencia a tomar vuelos largos durante la noche, se decidió subdividir en categorías tanto los horarios de partida de los vuelos, como la duración de los mismos. Finalmente obtuvimos dicho gráfico, del cual podemos concluir fácilmente que si bien tenemos gran cantidad de vuelos en la noche, no son la cantidad más grande dentro de los rangos.

Pero si nos ponemos a analizar la distribución de pasajeros específicamente en los vuelos nocturnos, podemos ver que es verdad que la mayoría de personas suele tomar vuelos largos, aunque no supere el 50% de los pasajeros. Aunque comparándolo con la cantidad de pasajeros que optan por realizar vuelos largos durante la mañana, es solo un poco más que la mitad.

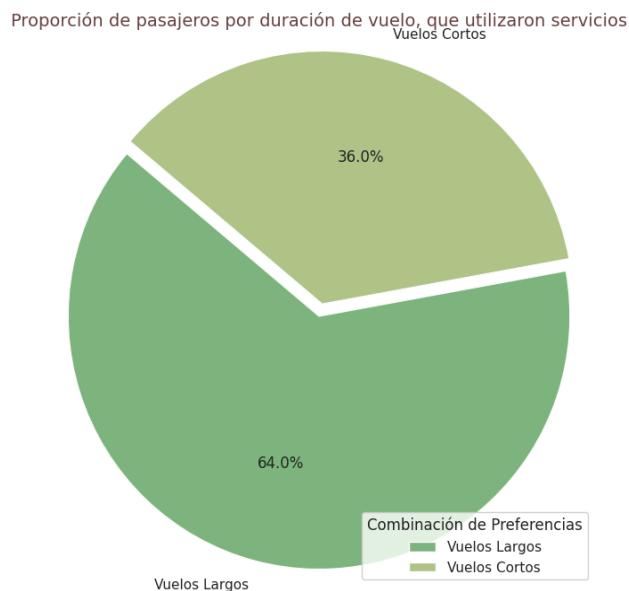


☞ [horario\\_duracion\\_pasajeros.ipynb](#)

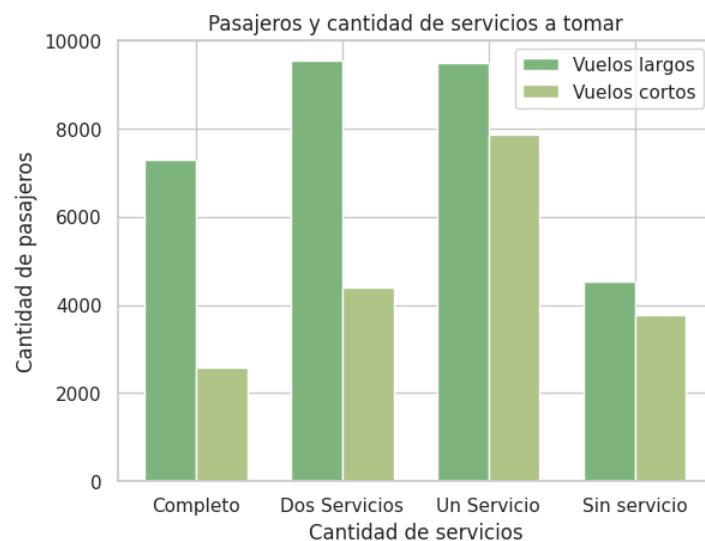
## Estudio tipo de vuelos y servicios

Al principio se supuso que la gente que toma vuelos largos, tiende a seleccionar servicios extras como: ***preferred\_seat, wants\_in\_flight\_meals, extra\_baggage***.

Para esto se tomó los vuelos comprados por Internet y se comparó con los vuelos cortos, teniendo en cuenta sólo a los que tomaron uno o más servicios. Podemos afirmar que hay más personas que toman servicios en vuelos largos.



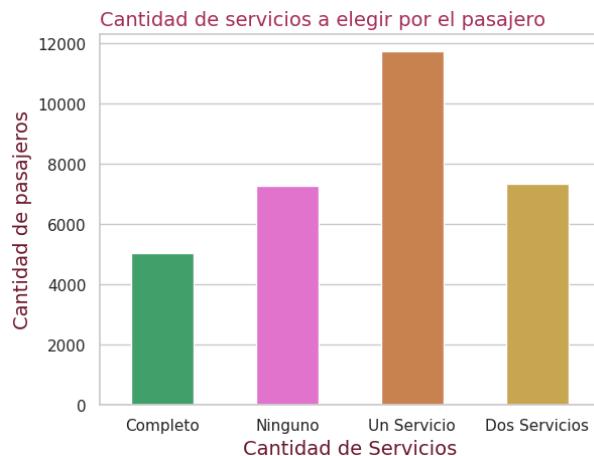
También podemos ver que tanto en vuelos largos como cortos, hay casi una misma cantidad de pasajeros que viajan sin servicios, y también podemos ver que dentro de los vuelos largos, la mayoría de los pasajeros viaja con uno o dos servicios, no es la mayoría la que viaja con el servicio completo.



☞ [pasajeros\\_servicios.ipynb](#)

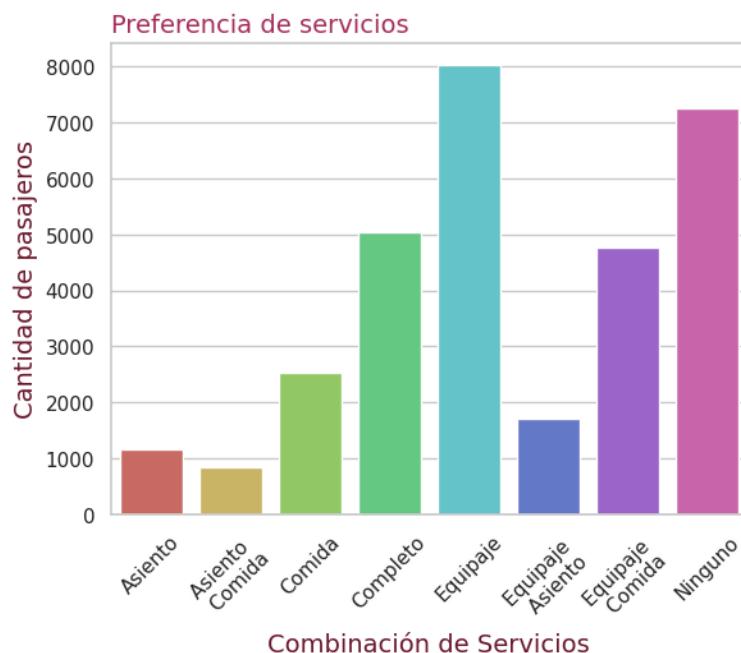
## Respecto a viajes individuales

A primera vista se había visto que la cantidad de vuelos individuales era mucho mayor a la cantidad de vuelos de 2 o más personas. Una suposición era que uno al viajar solo claramente prefiere viajar con todas las comodidades, es así que tenderían seguramente a seleccionar los servicios extras: **wants preferred seat, wants in flight meals, wants extra baggage.**



Entonces la cuestión fue cuántos eligen, y si había algún servicio en particular que eligen o alguna combinación de servicios favorita. Es así que del primer gráfico podemos decir, que la mayoría de los pasajeros tiende a elegir un servicio (aproximadamente el 37%), que el 24% prefiere combinar 2 servicios, el 23% viaja sin ninguno, y que es mínima la cantidad de personas que viajan con el servicio completo (aproximadamente el 16%).

Por otro lado viendo el segundo gráfico, notamos que el servicio favorito a la hora de elegir uno, es el **equipaje extra**, mientras que si hay que realizar una combinación las personas tienen a elegir **equipaje - comida**.



☞ [individuales\\_servicios\\_tendencia.ipynb](#)

## Visualización particular

¿CUANTOS COMENTARIOS TUVIMOS EN

# 2020?

Enero

Lu	Ma	Mi	Ju	Vi	Sa	Do
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	32	33	34	

Febrero

Lu	Ma	Mi	Ju	Vi	Sa	Do
				1	2	3
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29			

Marzo

Lu	Ma	Mi	Ju	Vi	Sa	Do
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

Abril

Lu	Ma	Mi	Ju	Vi	Sa	Do
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

Mayo

Lu	Ma	Mi	Ju	Vi	Sa	Do
			1	2	3	4
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

Junio

Lu	Ma	Mi	Ju	Vi	Sa	Do
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

Julio

Lu	Ma	Mi	Ju	Vi	Sa	Do
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Agosto

Lu	Ma	Mi	Ju	Vi	Sa	Do
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Septiembre

Lu	Ma	Mi	Ju	Vi	Sa	Do
					1	
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

Octubre

Lu	Ma	Mi	Ju	Vi	Sa	Do
1	2	3	4	5	6	
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

Noviembre

Lu	Ma	Mi	Ju	Vi	Sa	Do
			0	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

Diciembre

Lu	Ma	Mi	Ju	Vi	Sa	Do
					1	
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

cantidad de comentarios

0

10

20

30

40