Universidad Torcuato Di Tella

Examen domiciliario de R

Profesores: Bounos, Ian – Escobar, Martín

Alumno: Nicoliche, Leonardo Rafael

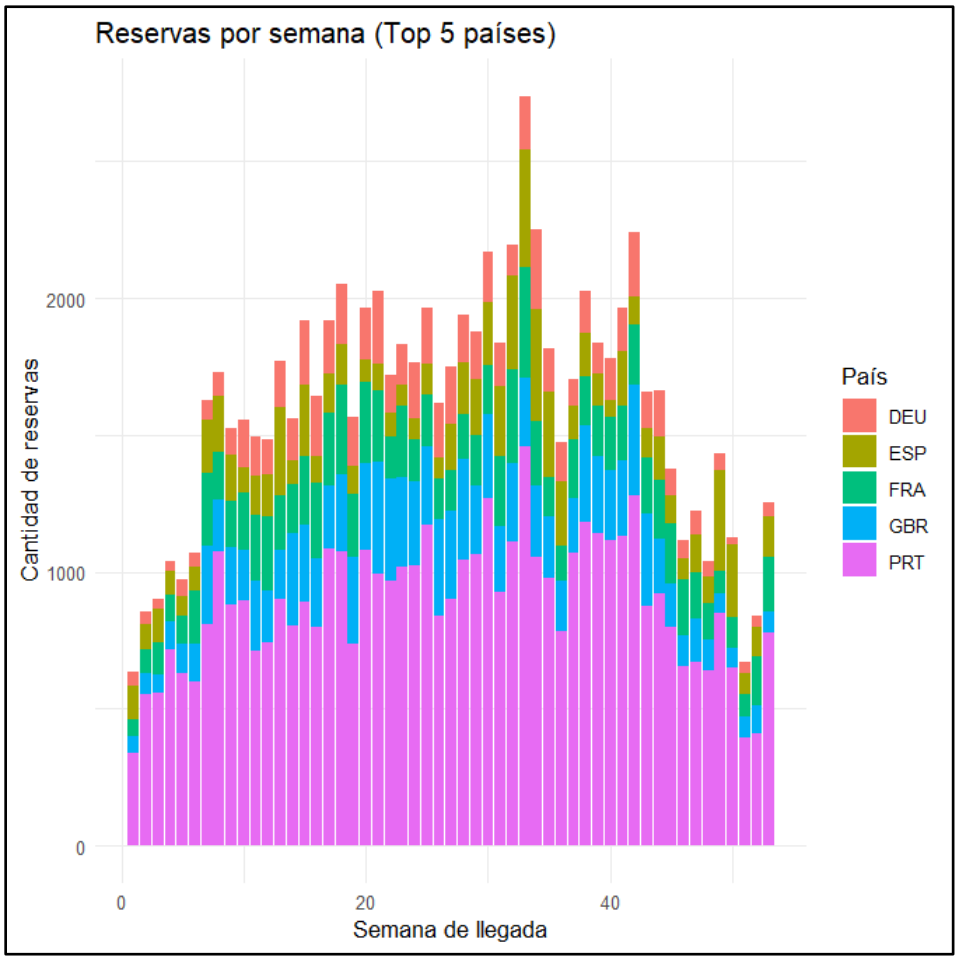
Parte 1

Se utiliza el dataset *hotel_bookings.csv* que contiene información sobre reservas hoteleras.

1.1

Pregunta 1: ¿Cómo se comportan las reservas por semana de los 5 países con más reservas?

Operando sobre el dataset original *datos_hoteles*, se contabilizó la cantidad total de reservas por país, identificando los cinco con mayor volumen. Luego se construyó el dataset *reservas_semana_paises*, que contiene las reservas semanales de esos países.



El gráfico de barras muestra que los cinco países mantienen un ratio relativamente constante entre sí, aunque se observa una estacionalidad: las reservas aumentan hacia la mitad del año y descienden hacia el final.

La tabla *top5_semanas* refuerza esta interpretación, ya que cuatro de las cinco semanas con mayor cantidad de reservas (30, 32, 33 y 34) se concentran en ese período casi consecutivamente.

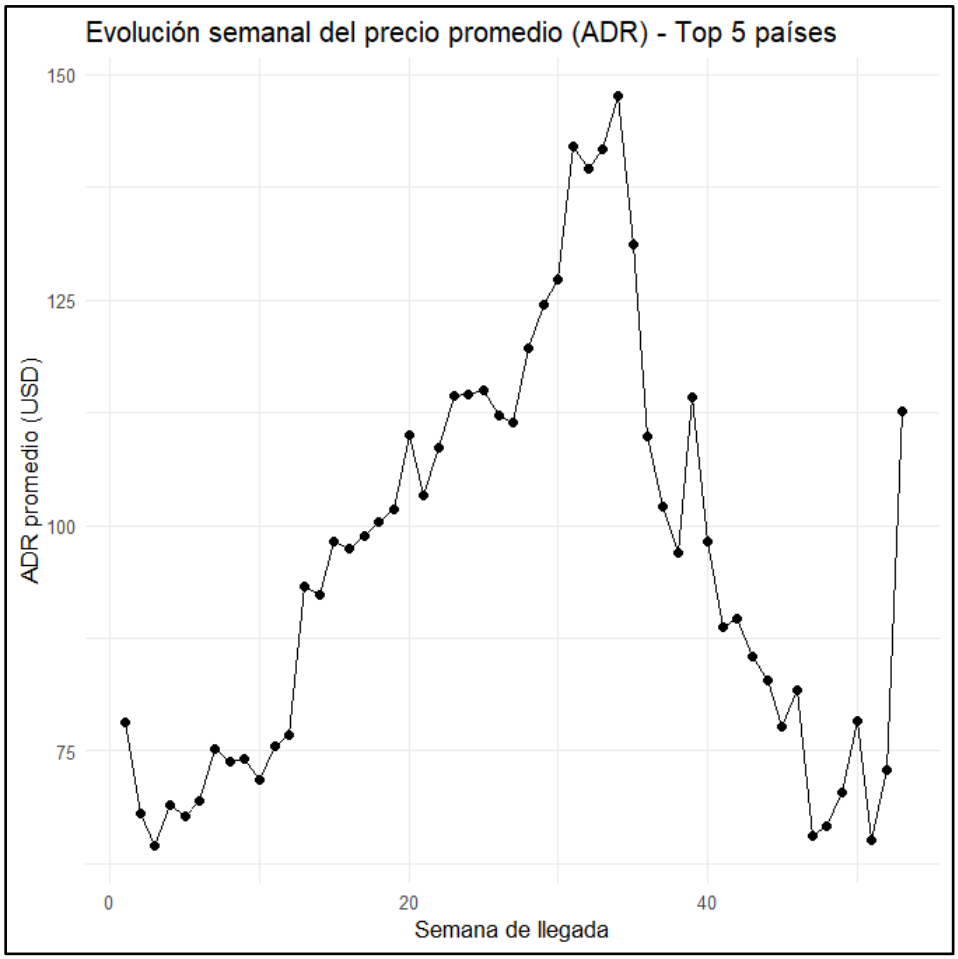
Arrival_date_week_number	Reservas_totales
33	2739
34	2252
42	2241
32	2197
30	2170

La siguiente pregunta busca indagar más al respecto.

1.2

Pregunta 2: ¿Cómo varía el ADR por semana para los 5 países con más reservas?

A partir del subconjunto de los cinco países con mayor cantidad de reservas, se calculó el promedio semanal del precio diario (*adr_por_semana*).



El gráfico muestra que el ADR aumenta progresivamente desde comienzos de año hasta

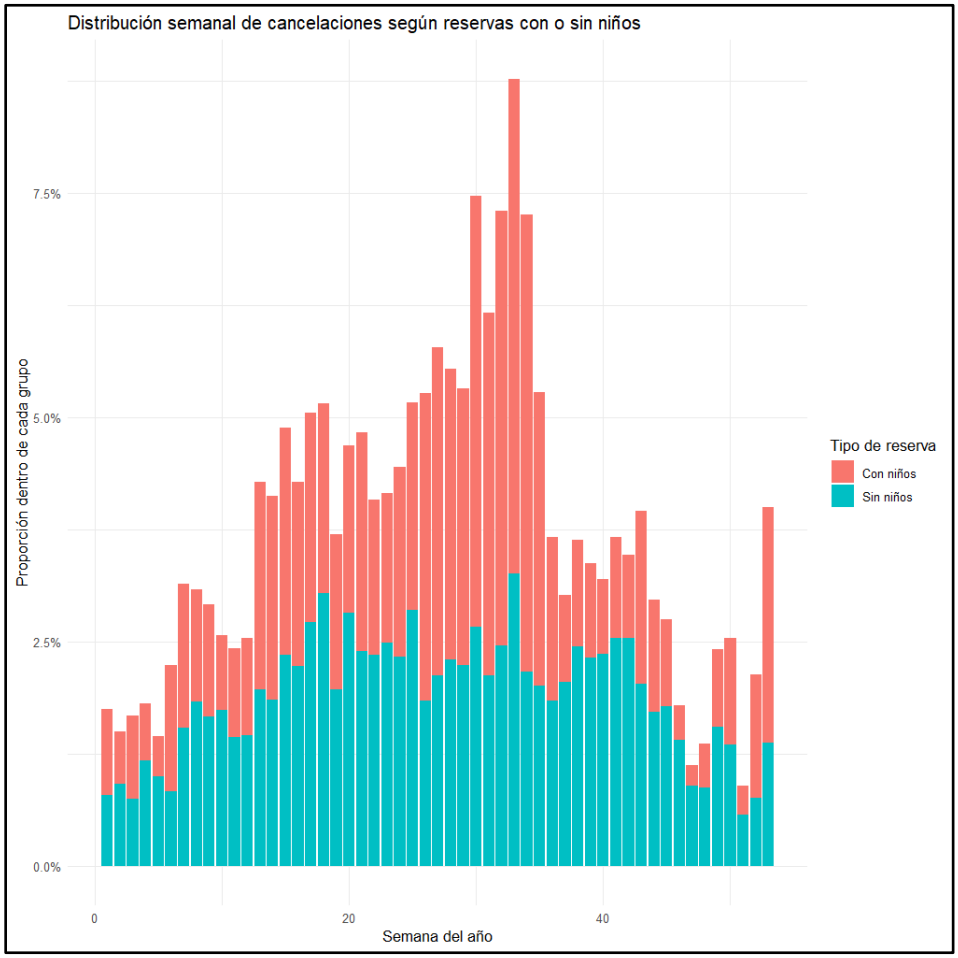
alrededor de la semana 30, alcanzando un máximo cercano a los 150 USD, y luego desciende con fuerza hacia fin de año.

Este comportamiento confirma una estacionalidad del precio, coherente con la observada en la cantidad de reservas: La demanda y el precio promedio se incrementan simultáneamente.

1.3

Pregunta 3: ¿Cómo se comportan las cancelaciones según la composición del grupo?

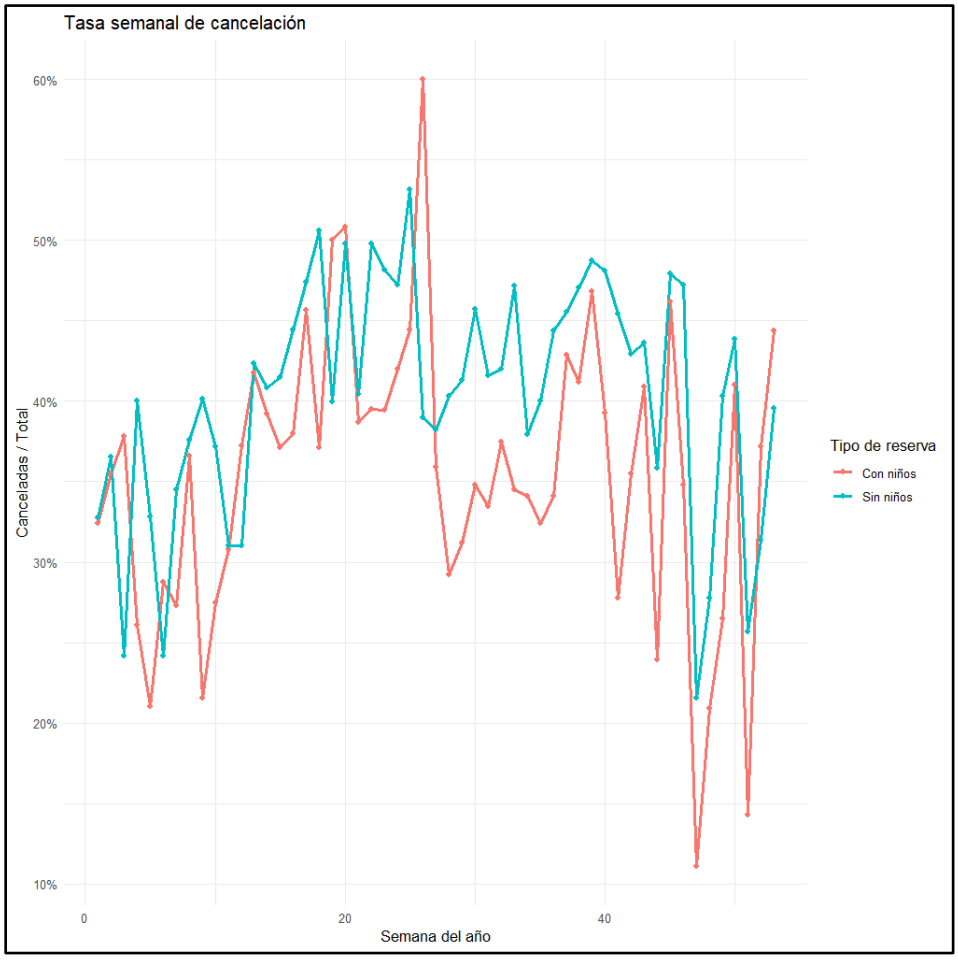
A partir del subconjunto de los cinco países con mayor cantidad de reservas, se filtraron únicamente las reservas canceladas y se clasificaron en dos categorías según la presencia de niños: "Con niños" y "Sin niños". *cancel_semana_comp* calcula, para cada semana, la proporción de cancelaciones dentro de cada grupo, de modo que cada serie ("Con niños" o "Sin niños") suma 1 a lo largo del año.



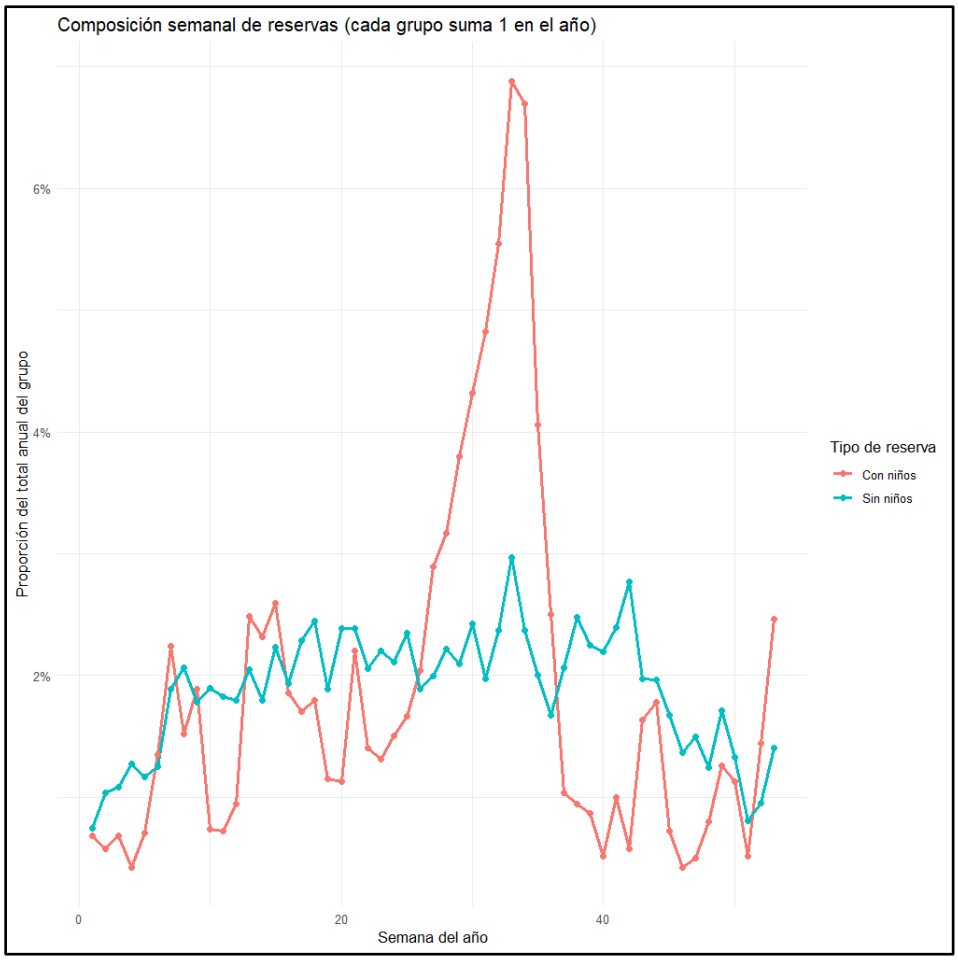
El gráfico muestra que las cancelaciones se concentran en las mismas semanas donde se observa el pico de reservas y de precios (ADR), alrededor de la semana 30. Es evidente cómo los grupos con niños son más elásticos ante la estacionalidad, la semana 20

equivale al 1.6% de las cancelaciones mientras que la semana 32 equivale al 6.62% (4.3 veces más).

Para interpretar este aumento se calculó desde la tasa de cancelación semanal según el grupo (canceladas / total de reservas).



El gráfico presenta una gran dispersión semanal y no muestra un aumento sostenido durante la temporada alta. Esto sugiere que el incremento de cancelaciones observado antes se debe principalmente a que hay más reservas totales en ese período, y no a que aumente la proporción de cancelaciones. Para corroborarlo se calculó la tasa de reservas semanal con la misma lógica.



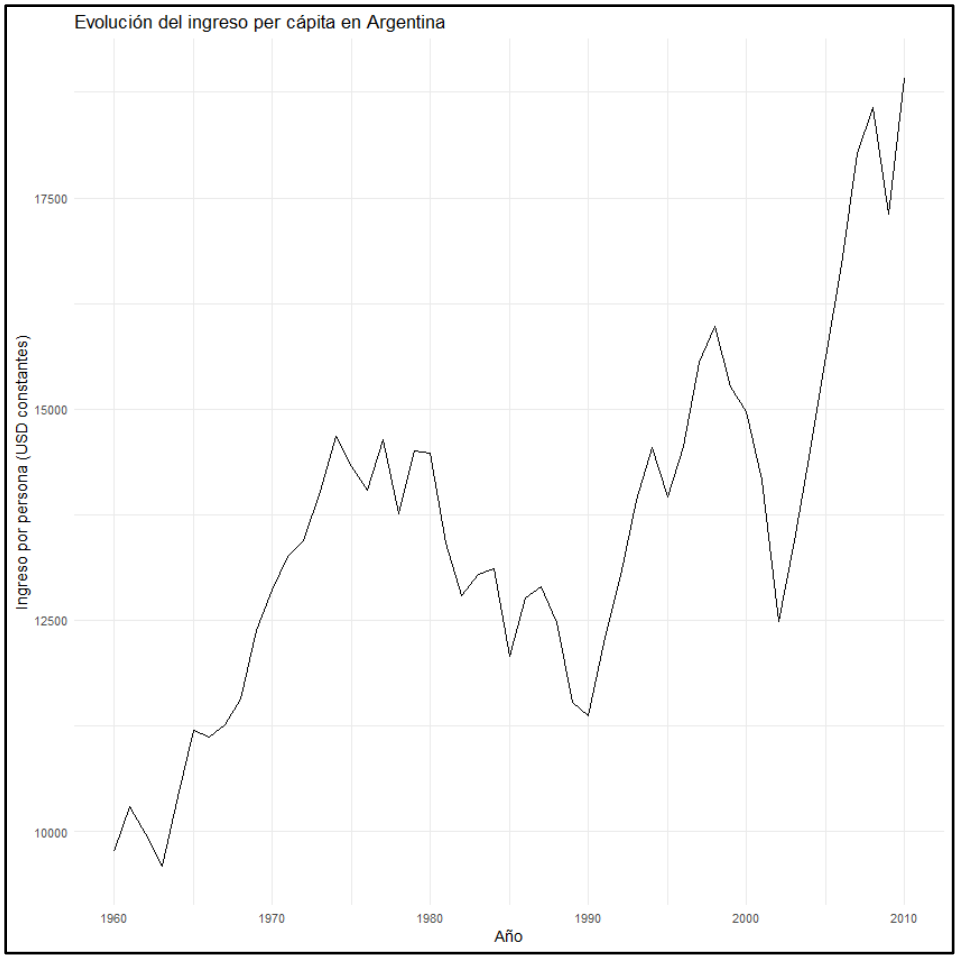
El spread (dispersión) observado en la distribución original de cancelaciones puede explicarse por un aumento desproporcionado de las reservas pertenecientes a grupos con niños durante la temporada alta. Aunque la tasa de cancelación no muestra un incremento sostenido, el volumen absoluto de reservas de este segmento crece significativamente en esas semanas, amplificando la cantidad total de cancelaciones observadas.

Conclusiones: Los datos reflejan una alta estacionalidad en las reservas hoteleras, que se traslada también a los precios —probablemente por fuerzas de mercado— y a las cancelaciones, aunque no por un cambio en el comportamiento de los huéspedes, sino por el aumento absoluto del número de reservas durante la temporada alta.

Parte 2

2.1

Utilizando el dataset de gapminder, se elaboró un gráfico de la evolución del ingreso per capita de Argentina a lo largo de los años.

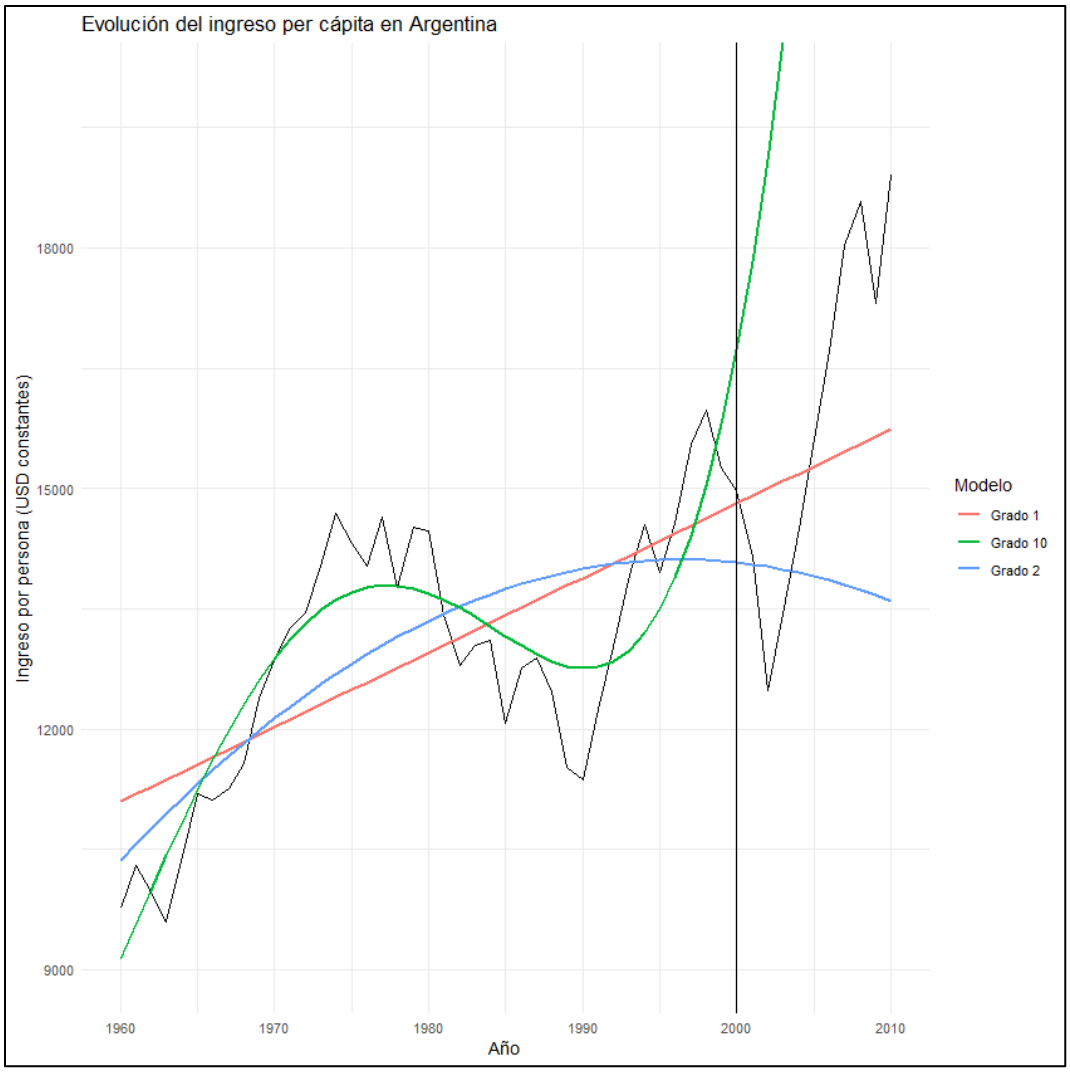


El gráfico muestra una tendencia creciente entre los años 1960 y 2010, a pesar de fluctuaciones en determinados periodos. Lo relevante, es que el valor correspondiente al año 2010 es sustancialmente mayor al de 1960.

2.2

Para el análisis, se dividió el conjunto de datos en dos subconjuntos: entrenamiento (1960–2000) y testeo (2001–2010). Sobre el conjunto de entrenamiento se ajustaron tres modelos distintos; un modelo lineal, un modelo polinómico de grado 2, y un modelo polinómico de grado 10.

Los resultados obtenidos se almacenaron en la tabla *grid_predicciones* y posteriormente se representaron gráficamente.



El gráfico muestra que, en el conjunto de entrenamiento, a medida que aumenta la complejidad del modelo, mejora la capacidad de ajuste a los datos. Sin embargo, en las predicciones sobre el conjunto de testeo, el modelo polinómico de grado 10, pese a presentar el mejor desempeño dentro de la muestra, sobreajusta (overfitting) y muestra un rendimiento deficiente fuera de la muestra (out-of-sample). Por el contrario, los modelos más simples, principalmente el modelo lineal, generan predicciones más verosímiles.

Esto puede ser evaluado con el RMSE in-sample y out-of-sample de cada modelo. Para ello se desarrolló una función que calcula el RMSE de cada modelo y guarda los resultados en dos tablas, dependiendo del conjunto de datos utilizado.

Modelo	Rmse_train
Grado 1	1212.74
Grado 2	1159.96
Grado 10	730.62

Modelo	Rmse_test
Grado 1	2012.47
Grado 2	3114.47
Grado 10	11311.21

Los resultados confirman la interpretación: el modelo de grado 10 presenta el menor RMSE in-sample, pero el mayor RMSE out-of-sample, mientras que el modelo lineal logra el menor RMSE fuera de muestra, demostrando una mejor capacidad de generalización.

2.3

Utilizando los datos correspondientes a Brasil, Chile, Uruguay y Perú, se construyeron tablas y calcularon las correlaciones con la función *cor()*. Adicionalmente para el crecimiento se crea una columna para medir la tasa de cada país.

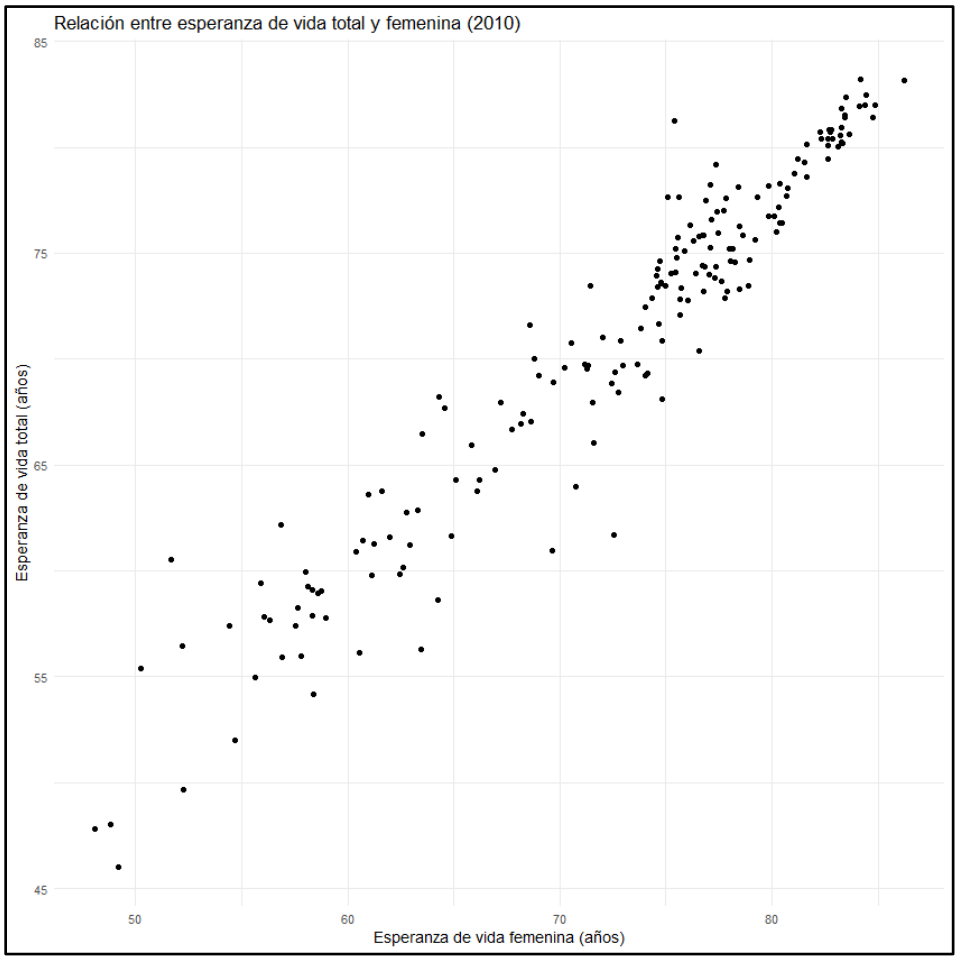
niveles_cor	Brasil	Chile	Perú	Uruguay
Brasil	1	0.771	0.564	0.871
Chile	0.771	1	0.548	0.940
Perú	0.564	0.548	1	0.577
Uruguay	0.871	0.940	0.577	1

crec_cor	Brasil	Chile	Perú	Uruguay
Brasil	1	0.006	0.410	0.277
Chile	0.006	1	0.043	0.365
Perú	0.410	0.043	1	0.412
Uruguay	0.277	0.365	0.412	1

Se pueden observar correlaciones positivas en los niveles de ingresos en los países bastante altas. Chile y Uruguay tiene 0.94 o Uruguay con Brazil 0.871. Por su parte los crecimientos al ser más volátiles no están tan fuertemente correlacionados. 0.365 y 0.277 son las correlaciones de los mismos países.

2.5

Utilizando el año 2010 se limpian los datos y se gráfica *life_expectancy* contra *life_expectancy_female*.



Se puede observar una clara correlación positiva casi de 1. Esto se debe a que contienen la misma información sobre la tendencia de la esperanza de vida, solo que una variable representa a un subconjunto de la otra.

2.6

El coeficiente estimado es de 0.89, por cada año que aumenta la esperanza de vida femenina lo hará en 0.89 la total. El R^2 es de 0.93, lo que significa que gran parte del comportamiento de la variable dependiente es explicable desde la variable independiente. Esto se debe a que *life_expectancy* es una combinación directa de *life_expectancy_female* con *life_expectancy_male.*

2.7

Al extraer el coeficiente y su error estándar, se calcula el estadístico t para contrastar si ese coeficiente es igual a 1, y se obtiene el intervalo de confianza al 95 %.

El resultado muestra que B1 = 0,886 con un t = −6,15, claramente menor al valor crítico. El intervalo [0.85 ; 0.92] no incluye 1, por lo que se rechaza la hipótesis nula. Al mirar los datos directamente, se observa que en todos los países la esperanza de vida femenina es mayor que la masculina y también que la total.

2.8

El modelo muestra que, manteniendo constante el nivel de ingresos, un aumento de un año en la esperanza de vida femenina se asocia con un incremento de aproximadamente 0,85 años en la esperanza de vida total. Por lo que la relación entre ambas variables sigue siendo muy fuerte e ínfimamente menor que en el modelo de una única variable.

Por su parte, el coeficiente del ingreso per cápita es positivo y de magnitud muy pequeña: un aumento de 10.000 dólares en el ingreso promedio se asocia con alrededor de 0,3 años más de esperanza de vida, siendo así una relación positiva pero casi despreciable en términos numéricos.

El R^2 pasa de 0,926 en la regresión simple a 0,929 en la múltiple, es decir, el ingreso apenas mejora la capacidad explicativa del modelo. En conclusión, incluir la variable *income_per_person* en términos cuantitativos aporta muy poco al modelo.

2.9

El dataframe contiene diversas variables, muchas de las cuales presentan una relación cualitativa muy ligada a la esperanza de vida. A partir de las mismas variables se construyeron distintos modelos, los cuales fueron evaluados en función de su complejidad y del coeficiente de determinación.

Entre ellos, el modelo que utiliza únicamente las variables *income_per_person* y *child_mortality* obtuvo un R^2 de 0.788, número inferior a los modelos con *life_expectancy_female*.
Aun así, este resultado adquiere relevancia dado que el modelo logra explicar una proporción considerable de la variabilidad en la esperanza de vida sin incluir variables directamente relacionadas.

Parte 3

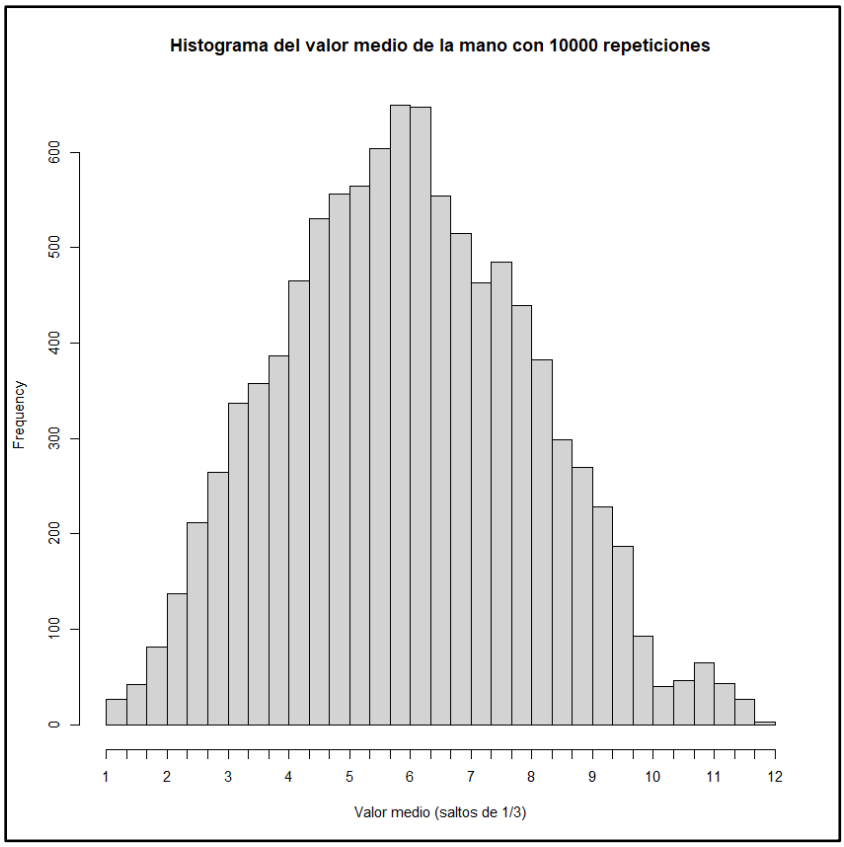
3.1

Para crear un mazo de truco se utilizó una tabla en donde se combinó los 4 palos y los 10 números factibles del mazo de truco en dos columnas, logrando que cada fila sea una de las cuarenta cartas.

3.2

Usando *sample()* se sortean 3 posiciones cualesquiera del mazo (cada una representando una de las cuarenta cartas), que se extraen a una nueva tabla. Del mazo original se retiran las 3 cartas de la mano usando *anti_join()*.

Se crea un vector *medias* con n elementos numéricos, donde, en un bucle *for* de n veces, cada iteración usará la función para simular una mano, calcularle la media y guardarla en el elemento número i. El histograma se calcula en base a este vector.



3.3

La función toma de input una mano aleatoria y le reemplaza los números mayores o iguales a 10 por 0, siendo este el puntaje que toman tales valores según las reglas. Luego crea una tabla auxiliar donde cuenta cuántas veces se repite en la mano cada palo y abre una condición con un *if*. Si acaso el palo que más se repite lo hace dos o más veces, volverá a la mano original y filtra por este mismo, sumando los dos valores más altos y sumándole 20. Si el palo más repetido aparece una única vez (3 palos distintos), buscará el valor más alto de la mano original.

3.4

Dada una mano aleatoria, la función utiliza 3 bucles, uno por cada posible carta del rival, para calcular todas las posibles combinaciones de las 37 cartas restantes (combinatorio 37C3). A cada mano posible le calcula los tantos con las mismas reglas y compara con los tantos de la mano original, clasificando el duelo para el rival como victoria o empate, aumentando el contador correspondiente en las variables *gana_oponente* y *empate*.

Por último, usando la regla de Laplace, se calcula la probabilidad de que el oponente gane dividiendo los casos favorables sobre los totales. Al discriminar si el oponente es mano o no, cambia quién gana en caso de empate, usando tales manos como caso favorable o no, según corresponda con las reglas.

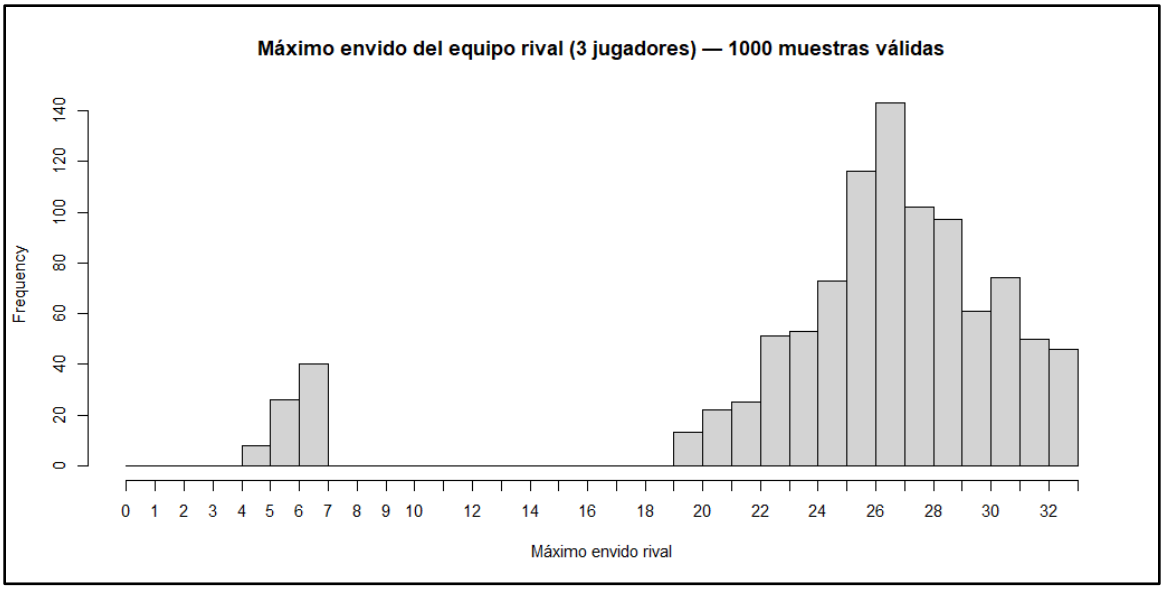
3.5

Calcular la probabilidad de ganar el envido utilizando la regla de Laplace conlleva mucha complejidad computacional, por lo que usaremos un estimador. La simulación arranca ya con la mano del jugador repartida. Se crean 5 manos al azar usando la misma lógica de *sample()* con las cartas disponibles y se reparten en grupos de a 3. Los dos primeros grupos de cartas son considerados de tus compañeros; para que este sorteo sea tenido en cuenta, se evalúa si ambos cumplen con la condición de su envido. Si es el caso, se suma al contador aceptadas, se calculan los tantos de los rivales y, siguiendo las reglas de truco, se decide qué equipo ganó. Si las manos de tus compañeros no cumplen con las condiciones, el sorteo no es considerado válido y no suma a ningún contador. El bucle se va a repetir hasta que se hayan hecho n sorteos válidos, definidos anteriormente en la variable *n_aceptadas_obj*.

Usando las simulaciones, se consigue un estimador insesgado y consistente para la probabilidad de ganar, calculado como manos ganadas / manos aceptadas.

3.6

Usando el mismo código pero con un vector que guarda los máximos envidos, dado una mano aleatoria (7 de basto, 3 de copa, 11 de espada) se construye el siguiente histograma.



Observando el gráfico, podemos ver cómo priman valores altos de envido para los rivales, porque ya no se trata de una única mano sino del máximo de 3 manos aleatorias. En consecuencia, la probabilidad de ganar los envidos se ve mucho más reducida que antes.