Ejercicio 1:

Para este ejercicio, como amantes del fútbol elegimos un dataset relacionado: resultados de fútbol desde 1872 hasta el presente 2025 (más específicamente de todos los dataset que tiene este más grande en Kaggle, usamos el de “results”). Comenzamos cargando las librerías y arreglando el formato de la fecha. Además verificamos que en los datos no tengamos ningún NAs.

Ahora que ya hicimos los chequeos preliminares, vemos como es y que contiene nuestro dataset. Vemos que este tiene datos de partidos, y contiene cosas como: qué tipo de torneo fue el que se jugó (amistoso, mundial, etc), que equipos jugaron y si jugaron de visitante o de local y cuantos goles hicieron. Vamos a pedirle ahora a R que cree una columna nueva que nos va a servir para nuestro análisis: le pedimos que esta columna se llame total_goals y sea la suma de los goles de visitante (away_goals) con los goles de local (home_goals).

Realizamos un gráfico preliminar que nos dice cuántas veces aparece una X cantidad de goles totales: por ejemplo, vemos que los partidos más comunes son los que tienen uno o dos goles en total, mientras que los más raros son los que tienen más de cinco.

Ahora vamos con nuestra primera pregunta: ¿Cómo evolucionó el promedio de goles por partido por década?. Lo que nos preguntamos es en qué época se hacían más goles, si con el fútbol de “antes” o con el fútbol “moderno”. Para eso, creamos una columna que agrupe los datos por década y le pedimos a R que calcule dos cosas: el número de partidos en esa década y el promedio de goles por partido en en la misma. Luego de esto, hacemos un gráfico y podemos ver que antes se

hacían en promedio muchos más goles que ahora. Esto muestra la evolución del fútbol como deporte y como se fue profesionalizando, ya que ahora es más difícil hacer goles debido a los avances en táctica y técnica así como en entrenamiento

(tanto salud mental como salud física).

La siguiente pregunta qué queremos responder es: ¿Ser local te da más ventaja de goles? Esto es algo que en el fútbol se suele decir mucho, si jugas de local siempre se cree que tenes mas chances de ganar o de sacarle ventaja al equipo rival. Pero veamos si es así y si cambia con el tiempo. Ahora le vamos a pedir a R calcular la diferencia de goles promedio por década, esto lo vamos a hacer haciendo los goles de local menos los goles de visitante. Al hacer el gráfico vemos que en promedio, la diferencia entre los goles que un equipo hace de visitante y los goles que hace de local suele ser positiva y que esta además se mantuvo positiva a lo largo de las décadas. Esto responde y confirma lo que dijimos al principio: el equipo local suele tener las de ganar. Esto puede deberse a varios factores: presión de la hinchada rival, "jet lag" por el viaje en caso de que sea un partido internacional y hasta la ubicación del estadio puede afectar (véase los partidos en la altura, los jugadores locales suelen estar más acostumbrados mientras que los visitantes suelen sufrir bastante. Esto hace que sea mucho más probable que los locales saquen ventaja.)

Finalmente la última pregunta que responderemos es: ¿Hay torneos donde se hacen más goles que otros? Básicamente: ¿hay alguna competición donde se suelen convertir más goles que en otra? Para responder esto, primero vamos a ver los diez torneos más comunes en nuestro dataset (o sea los diez torneos con más partidos registrados). Al graficar estos datos para cada torneo vemos que, aunque

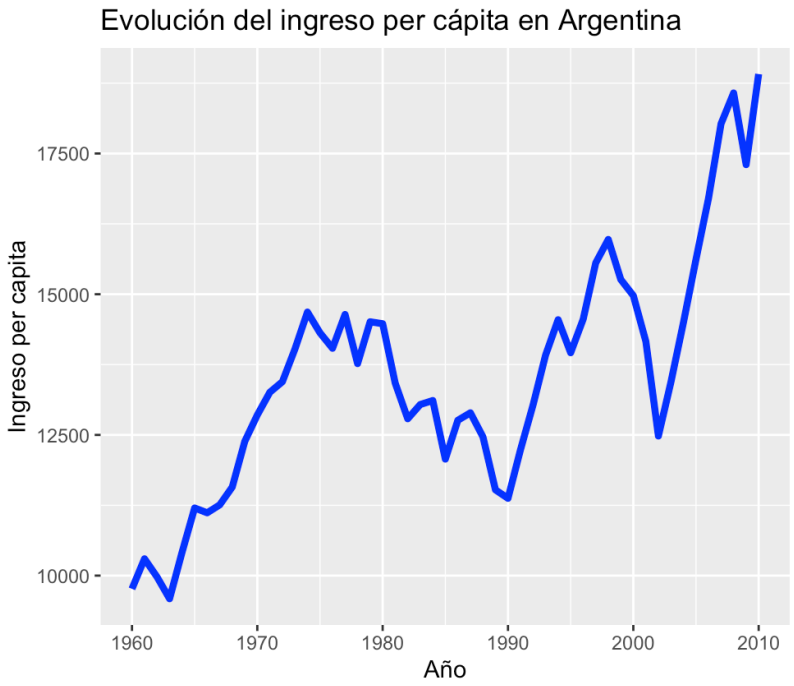
algunos tienen más fluctuaciones que otros y más "casos extremos" (véase por ejemplo las eliminatorias del mundial), la mayor parte de las distribuciones está concentrada entre cero y dos goles. Esto nos dice entonces, que no hay torneos donde se hagan más goles que otros.

Ejercicio 2:

Parte 1: Ingreso por persona

<u>Inciso 1:</u>

Primero bajamos la base de datos de *Gapminder* y después filtramos el dataset para quedarnos solo con los datos de Argentina. Los guardamos en una nueva variable y armamos un gráfico que muestra cómo fue cambiando el ingreso por persona a lo largo del tiempo. El resultado fue este:



Desde los 60 se ve una tendencia de crecimiento hasta mediados de los 70. Después se nota una caída fuerte, que coincide con la etapa de hiperinflación del gobierno de Alfonsín. Más adelante hay una recuperación durante los años de la convertibilidad, cuando los ingresos vuelven a subir y llegan un poco por encima del nivel previo a la dictadura. En los 2000 vuelve a caer de forma brusca, justo en el momento del corralito y la crisis de 2001. Finalmente, se ve una nueva recuperación con el gobierno de Néstor Kirchner, que sigue en los primeros años del gobierno de Cristina. Pero a partir de 2010 la curva se aplana y aparece el estancamiento que muchos llaman “la década perdida”.

<u>Inciso 2:</u>

Para este inciso elegimos los datos de Argentina y dividimos el conjunto en dos partes:

- **Entrenamiento (train):** incluye todos los años menos los últimos 10.

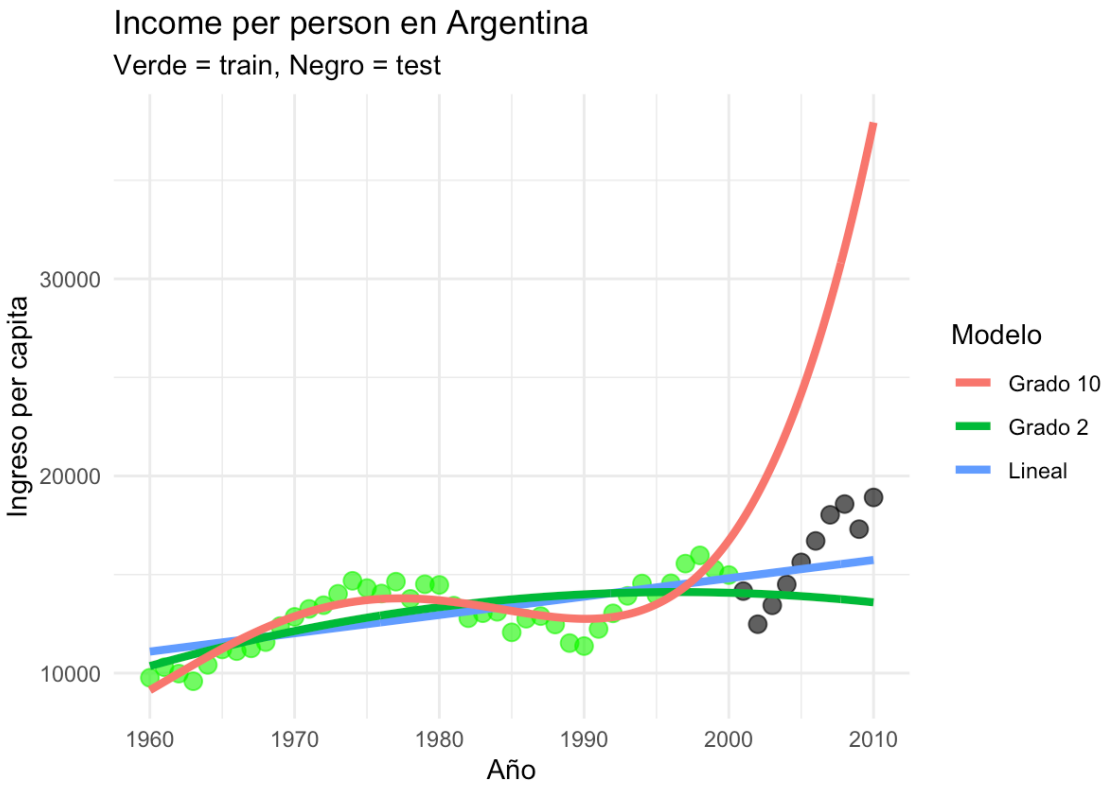
- **Testeo (test):** incluye justamente esos últimos 10 años.

La idea es usar la parte de *test* para ver qué tan bien funcionan los modelos con datos que no se usaron al entrenar.

Después ajustamos tres modelos de regresión, usando el tiempo como variable independiente:

- Modelo lineal.
- Modelo polinómico de grado 2.
- Modelo polinómico de grado 10.

Después obtuvimos las predicciones de cada modelo y armamos un gráfico donde se ven los puntos reales (tanto del conjunto *train* como del *test*) junto con las líneas de las distintas regresiones. De esa manera podemos ver visualmente qué tan bien cada modelo logra acercarse a los datos reales.



Se puede ver que el modelo lineal no logra captar la forma curva de los datos, aunque resulta ser el más estable. En cambio, el modelo cuadrático (grado 2) se adapta mejor a la tendencia general, pero no representa bien los valores de los extremos. Finalmente, el modelo de grado 10 se ajusta casi perfecto al conjunto de

entrenamiento, pero exagera los cambios (lo que se conoce como *overfitting*) y no logra generalizar bien cuando lo probamos con los datos de test.

Esto se confirma cuando calculamos la raíz del error cuadrático medio (RMSE), que mide la diferencia entre los valores reales y las predicciones. Cuanto más chico sea el RMSE, mejor es el modelo.
Calculamos el RMSE tanto para *train* como para *test* y obtuvimos lo siguiente:

```
RMSE en TRAIN:
   Lineal    Grado2   Grado10
1212.7475 1159.9654  730.6231
> cat("\nRMSE en TEST:\n"); print(rmse_test)

RMSE en TEST:
   Lineal    Grado2   Grado10
 2012.472  3114.474 11311.216
```

Se puede ver que en la muestra *train* el modelo de grado 10 tiene menor error que el resto. Esto empeora en la muestra *test*, por lo tanto podemos decir que hay overfitting.

Inciso 3:

A)

Creamos una matriz de correlación del ingreso per cápita de los países: Argentina, Colombia, Chile, Bolivia y Ecuador.

Para poder hacer esto filtramos el dataset original (el cual contiene los datos de todos los países) por los países elegidos (los que nombramos antes).

La matriz da como resultado correlaciones positivas y altas entre países lo cual significa que en el largo plazo los niveles de ingreso entre estos países es similar.

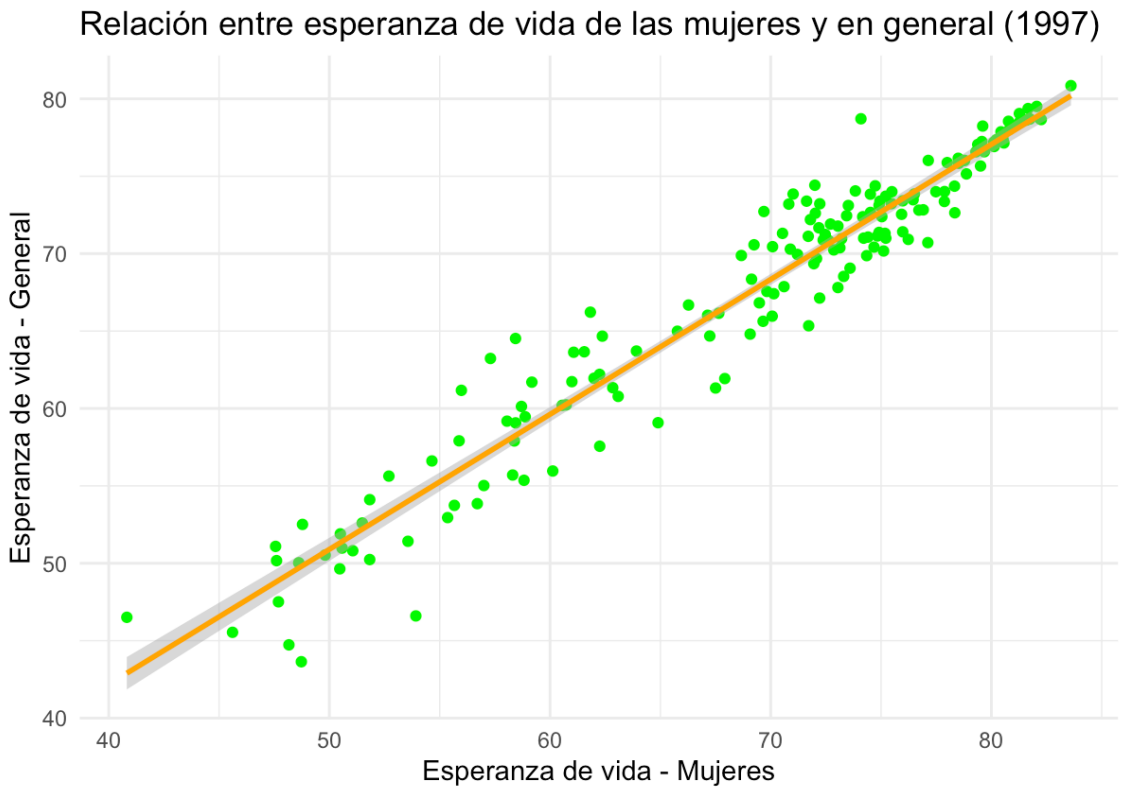
B) En este inciso comparamos las variaciones porcentuales anuales. Vemos que la matriz de correlaciones da como resultado números mucho mas bajos que la del inciso anterior. Esto significa que aunque es verdad que en el largo plazo los países terminan teniendo niveles de ingreso muy parecidos, si medimos año por año los países varían mucho entre sí ya que tienen diferentes factores afectandolos.

Parte 2: Esperanza de vida y género

Inciso 5:

Para esta segunda parte elegimos analizar el año 1997, por lo tanto filtramos el dataset original para que nos aparezcan solo los datos de ese año.

Hacemos el siguiente gráfico:



El gráfico muestra una relación positiva y lineal entre la esperanza de vida de las mujeres y la esperanza de vida en general en el año 1997. Lo cual es algo intuitivo y lógico ya que si aumenta la esperanza de vida de las mujeres (un sector de la sociedad) consecuentemente va a aumentar la esperanza de vida de la sociedad en general.

Inciso 6:

Hacemos una regresión lineal simple entre la esperanza de vida de las mujeres y la esperanza de vida en general.

El dato que obtuvimos del coeficiente de esperanza de vida de las mujeres es alto lo cual tiene sentido ya que como explicamos antes hay una correlación fuerte entre la esperanza de vida en general y la esperanza de vida de la sociedad.

También podemos observar el dato del R2 el cual también da un número alto. Esto demuestra lo que veníamos diciendo, la esperanza de vida de las mujeres tiene mucha influencia sobre el dato de la esperanza de vida en general.

Inciso 7:

Planteamos una hipótesis nula (H0) la cual plantea que la esperanza de vida general = la esperanza de vida de las mujeres y una hipótesis alternativa (H1) la cual plantea que la esperanza de vida de las mujeres es mayoir a la esperanza de vida en general. Hacemos un test de T y obtenemos un p-valor muy bajo (casi 0) por lo tanto para un test con nivel de significancia del 5% podemos afirmar que se rechaza H0. Es decir que la esperanza de vida de las mujeres es mayor que la esperanza de vida general.

Inciso 8:

Hacemos una regresión múltiple. estimamos esperanza de vida general sobre esperanza de vida de las mujeres y el ingreso por persona.

Si comparamos los resultados obtenidos en los Test con y sin el ingreso por persona vemos los siguientes resultados:

SIN EL INGRESO POR PERSONA:

```
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             7.30613    1.23439   5.919 1.82e-08 ***
life_expectancy_female  0.87190    0.01772  49.192  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.298 on 165 degrees of freedom
Multiple R-squared:  0.9362,    Adjusted R-squared:  0.9358
F-statistic:  2420 on 1 and 165 DF,  p-value: < 2.2e-16
```

CON INGRESO POR PERSONA:

```
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           8.041e+00  1.370e+00   5.871 2.34e-08 ***
life_expectancy_female 8.584e-01  2.084e-02  41.186  < 2e-16 ***
income_per_person      1.455e-05  1.183e-05   1.230     0.22
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.295 on 164 degrees of freedom
Multiple R-squared:  0.9368,    Adjusted R-squared:  0.936
F-statistic:  1214 on 2 and 164 DF,  p-value: < 2.2e-16
```

Se puede ver que la esperanza de vida de las mujeres cae un poco cuando se tiene en cuenta el ingreso por persona pero sigue siendo un valor muy alto el cual tiene mucho impacto en el dato de la esperanza de vida en general.

Vemos que el valor del coeficiente de ingreso por persona es muy bajo lo cual significa que este es un dato el cual es muy poco relevante para la esperanza de vida general.

El dato del R2 aumenta (no mucho) del modelo simple al modelo múltiple. Lo cual nos da a entender que la adición del ingreso por persona mejora un poco la estimación de la esperanza de vida general. (Hay que tener en cuenta que estos son datos tomados solo del año 1997).

Inciso 9:

En este inciso elegimos utilizar las variables: ingreso por persona y esperanza de vida de los hombres para explicar la esperanza de vida general. (Aunque anteriormente dijimos que el ingreso por persona no es muy relevante a la esperanza de vida elegimos incluirlo en este inciso ya que vimos que cuando se usa este dato el R2 aumento por lo tanto este dato puede llegar a afectar a la esperanza de vida total más de lo esperado).

Esta regresión nos dio como resultado los siguientes datos:

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.237e+00  1.433e+00   4.352 2.36e-05 ***
life_expectancy_male  9.609e-01  2.366e-02  40.605  < 2e-16 ***
income_per_person    -1.360e-05  1.236e-05  -1.100    0.273
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.325 on 164 degrees of freedom
Multiple R-squared:  0.9351,    Adjusted R-squared:  0.9343
F-statistic:  1181 on 2 and 164 DF,  p-value: < 2.2e-16
```

Vemos primero que la esperanza de vida de los hombres da un valor muy alto. Esto significa que esta variable va a afectar mucho a la esperanza de vida general (por lo tanto es una buena variable explicativa).

La segunda variable utilizada es ingreso por persona, la cual dio un valor menor a 0. Esto significa que a mayor ingreso por persona menor esperanza de vida general va a haber. Igualmente podemos observar que el valor estimado de esta variable es muy chico por lo tanto entendemos que esta variable finalmente no es una buena variable explicativa para analizar la esperanza de vida general.

Finalmente vemos que el valor del R2 es muy alto lo cual significa que la variable esperanza de vida general es explicada en gran parte por las variables incluidas en este modelo.

Ejercicio 3:

1) Debemos armar una función a la que si le damos ciertos datos, nos devuelva la cantidad de barcos perdidos tanto por parte del atacante como del defensor en una ronda siguiendo la lógica del juego de la batalla naval descripta en el código. Acá un ejemplo con 3 barcos para el atacante y 2 para el defensor

```
> #Pruebo los comandos con datos:
> # Ejecutar una comparación con 3 dados atacante y 2 del defensor
> resultado_ataque(3, 2)
perdidas_atacante perdidas_defensor
                1                 1
> # Resultado esperado: vector c(perdidas_atacante = X, perdidas_defensor = Y)
```

2) En este caso, tendremos que simular una batalla completa. La batalla finaliza cuando el atacante sea victorioso, o de lo contrario, cuando se quede con una sola ficha. Si el atacante conquista, obtendremos un comando respuesta "TRUE". Acá una simulación de (5,3):

```
> # Simular batalla empezando con 5 fichas atacante y 3 defensor
> simular_batalla(5, 3)
$fichas_atacante_final
perdidas_atacante
                5

$fichas_defensor_final
perdidas_defensor
                0

$atacante_conquista
perdidas_defensor
             TRUE
```

Podemos ver que, al quedarse el defensor sin fichas, efectivamente nos arroja "TRUE". Sin embargo, no existe argumento para indicar que de correr nuevamente la simulación no nos diga "FALSE", es aleatorio.

3) Ahora se nos plantea que simulemos 1000 batallas en las que siempre cada lado tenga 5 buques, el código que planteamos nos da el siguiente resultado:

```
> # Estimación precisa (10000 simulaciones)
> probabilidad_ataque(10000, 5, 5)
[1] 0.191
```

Corriendo en reiteradas ocasiones el código, nos arroja siempre una probabilidad que ronda el 19%, como vemos en esta captura extraída del archivo del ejercicio 3, lo que nos lleva a intuir que si tomamos una muestra lo suficientemente grande, ante igualdad de barcos por cada parte, existe una cantidad considerablemente menor de que el territorio sea conquistado a que se produzca lo contrario.