|  |  |
| --- | --- |
| Fundamentos de la Analítica de Datos | |
| About | Ciencias de la Computación Segundo Semestre UTEM |  |
| PROFESORA: DIANA LÓPEZ NOMBRE:CAMILA CARRASCO LATIN FECHA: 29/04/2022 |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
| INTRODUCCIÓN | | |

|  |  |
| --- | --- |
| Software de estudio Para el desarrollo del módulo I: Fundamentos de la Analitica de Datos, utilizaremos la herramienta analitica R- studio. Este software corresponde a un entorno y lenguaje de programación orientado al análisis estadístico, al cálculo, manipulación de datos, y representaciones gráficas.  Para el desarrollo de este trabajo utilizaremos Rstudio con la instalación de distintos paquetes y librerías.  Las ventajas de este IDE es que nos permitirá abrir varios scripts a la vez, ejecutar pedazos de código con solo marcarlo y es compatible con todos los sistemas operativos. |  |
|  | |  |
| |  |  | | --- | --- | |  | Descripción El análisis presentado en este informe pretende abordar la salud física de las personas de acuerdo a una serie de factores.  Es sabido que el descanso es una necesidad del organismo, es por esto que una cantidad inadecuada repercute directamente sobre el bienestar de una persona.  De acuerdo con la Asociación Mundial de Medicina del Sueño (WASM, por sus siglas en inglés),"los problemas de sueño constituyen una epidemia global que amenaza la salud y la calidad de vida de más del 45% de la población mundial", por lo tanto dormir bien es un pilar fundamentales para gozar de una buena salud, además de una dieta equilibrada y ejercicio regular.  Quedarse en la cama sin poder dormir y otros errores que empeoran el  insomnio | | |  |

|  |  |  |
| --- | --- | --- |
| En primera instancia asociamos como problema la falta de sueño por su relación con problemas físicos , por ejemplo tenemos el riesgo de presión arterial alto. En lo que respecta al corazón se produce una disminución de la temperatura corporal basal, responsable entre otros factores de la constancia del ritmo cardiaco. Otra enfermedad asociada es la diabetes, ya que la falta de sueño repercute en la habilidad de procesar la glucosa y esto provoca altos niveles de azúcar.  Finalmente tenemos el accidente cerebrovascular.  Se han demostrado otras enfermedades a largo plazo, como la hiperalgesia, o aumento de la percepción del dolor, los temblores, el envejecimiento precoz, el agotamiento y los trastornos gastrointestinales.  En el caso cuando la cantidad de horas supera las 8 horas también nos encontramos en un escenario riesgoso, ya que estamos propensos a sufrir otras o las mismas enfermedades. El sueño prolongado produce alteraciones metabólicas por qué hacemos menos actividad física y comemos fuera de horario. En lo que respecta a enfermedades el sueño prolongado también está relacionado estrechamente con accidentes cardiovasculares e infartos.  La diabetes podría catalogarse como un riesgo, ya que dormir demasiado eleva los niveles de azúcar, al igual que cuando se duerme muy poco.  Las características mencionadas anteriormente se relacionan a morir prematuramente, estudios aseguran que las personas que  duermen demasiado mueren prematuramente, esto por que la probabilidad de morir aumenta padeciendo diabetes o alguna enfermedad cardiovascular.  Nombraremos algunos casos donde el tiempo de sueño puede ser prolongado:   * Narcolepsia: Asociado a sueño profundo de manera repentina, tiene relación con la debilidad muscular parcial. * Hipersomnia idiopática: Necesidad de alargar las horas de sueño, sin causa conocida. El cuerpo te pide entre 12, 14 y 16 horas de sueño. * Síndrome de Kleine y Levin: Afecta normalmente a jóvenes varones y se traduce en excesos de 16 o 18 horas de sueño diario. | | |
|  |  | Dataset El dataset utilizado en este informe pertenece a la plataforma web Kaggle, la cual reúne la comunidad Data Science más grande del mundo. Kaggle permite acceder de manera gratuita a una gran cantidad de datos publicados por la comunidad.  El dataset lleva por nombre indicadores clave personales de enfermedades cardiacas. Estos datos son de la encuesta anual realizada por CDC en el año 2020 (Centros para el Control y Prevención de Enfermedades). El link del dataset será adjunto a continuación:  https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/metadata  **Origen de los datos**  El conjunto original de datos fue obtenido de las encuestas que se realizan de manera telefónica y son realizadas cada año por la CDC. La información que obtenemos es que inicialmente este conjunto contenía 401.958 filas y 279 columnas, las cuales fueron reducidas a datos de interés quedando 319.795 filas y 18 columnas. Este conjunto fue tratado para evaluar los factores que afectan o pueden ser causantes de un paro cardiaco.  En nuestro caso, pretendemos usar esta misma información pero para determinar la correlación entre variables, en particular dar respuesta a la afirmación que establece que las las horas de sueño no están relacionadas con la cantidad de días que una persona tiene deficit de salud fisica. |  |  |

### Variables

Las variables de las columnas fueron obtenidas a partir del Informe del libro de códigos LLCP 2020. Este Libro contiene el nombre de la variable, la ubicación y la frecuencia de los valores para todas las áreas de informes combinadas para el conjunto combinado de datos.

El dataset fue creado con las variables que se estimo relevantes en este estudio. En el siguiente enlace se puede observar todas las variables que formaban parte de este estudio.

https://www.cdc.gov/brfss/annual\_data/2020/pdf/codebook20\_llcp-v2-508.pdf

* **HeartDisease:** Esta variable corresponde a si el consultado alguna vez diagnosticado con ataque al corazón
* **BMI:** Índice de masa corporal calculado. Con el sistema métrico, la fórmula para el IMC es el peso en kilogramos dividido por la estatura en metros cuadrados.
* **Smoking:** Estado de tabaquismo en la actualidad, en estado fumador y no fumador
* **AlcoholDrinking:** Estado de tabaquismo en la actualidad, en estado bebedor y no bebedor
* **Stroke:** Esta variable corresponde a si el consultado alguna vez diagnosticado con accidente cerebrovascular
* **PhysicalHealth:** Número de días que la salud física no es buena, incluye el no poder realizar actividades habituales, como el cuidado personal, el trabajo o la recreación.
* **MentalHealth:** Número de días que la salud mental no es buena, incluye estrés, depresión y problemas con las emociones.
* **DiffWalking:** Dificultad para caminar o subir escaleras
* **Sex:** Sexo del encuestado.
* **AgeCategory:** Edad reportada en categorías de edad de cinco años.
* **Race:** valor de raza/etnicidad imputada, si la o el encuestado se negó a dar una raza/etnicidad. El valor de la raza/etnicidad imputada será la más común respuesta de raza/etnicidad para esa región del estado)
* **Diabetic:** Esta variable corresponde a si el consultado alguna vez diagnosticado con ataque al corazón
* **PhysicalActivity:** Adultos que informaron haber realizado actividad física o ejercicio durante los últimos 30 días además de su trabajo habitual.
* **GenHealth:** Salud general
* **SleepTime:** Cantidad de horas que duerme en un período de 24 horas
* **Asthma:** Esta variable corresponde a si el consultado alguna vez diagnosticado con asma.
* **KidneyDisease:** Esta variable corresponde a si el consultado alguna vez diagnosticado con enfermedad renal, Sin incluir cálculos renales, infección de la vejiga o incontinencia.
* **SkinCancer:** Esta variable corresponde a si el consultado alguna vez diagnosticado con cáncer de piel.

### Análisis estadístico

Planteamiento de las hipótesis de este estudio

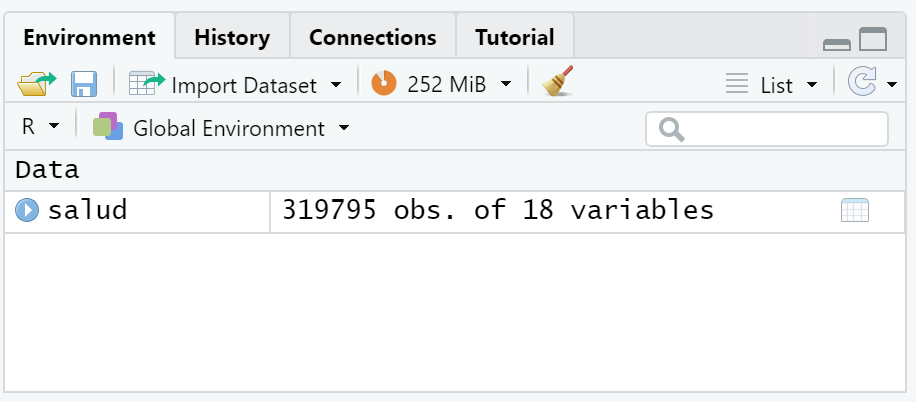
H0: No existe un grado de correlación entre las horas que duerme una persona y la cantidad de días que tendrá una salud física deficiente.

H1: La cantidad de horas que duerme una persona está relacionada con la cantidad de días que su salud física será deficiente.

Realizamos la carga del dataset a partir de nuestro carpeta de origen , a este conjunto de datos lo nombraremos salud



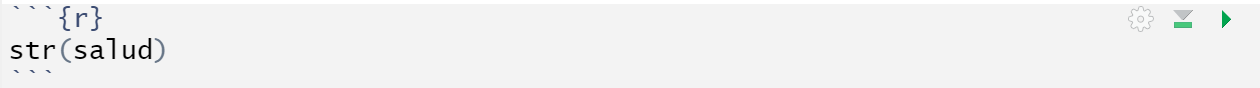
Como observamos nuestro dataset se cargó correctamente y se nos indica de inmediato la cantidad de observaciones que corresponden a las filas y 18 variables correspondientes a las columnas.



Si apretamos este conjunto se nos despliega una nueva ventana llamada salud, la cual nos permite tener una vista de nuestros datos.



Mostraremos las características de la base de datos, donde podemos observar que mayoritariamente tenemos variables chr. Sin embargo, nuestras variables de interés pertenecen al tipo num. También esta función nos indica la cantidad de observaciones y de variables.

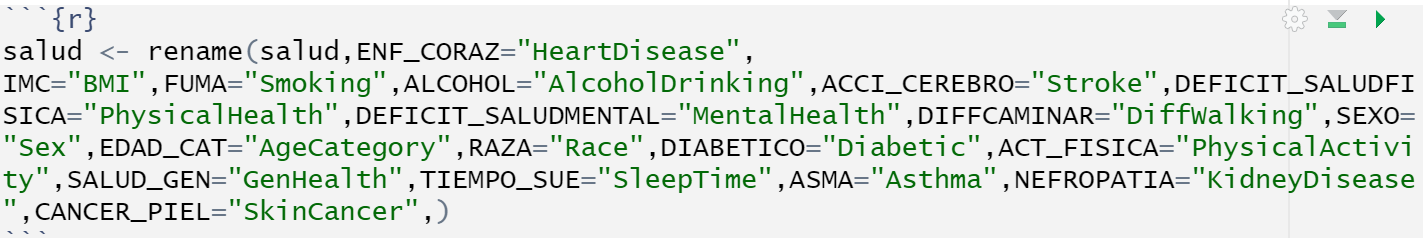


Texto, Carta

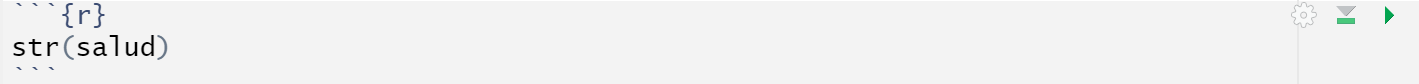
Descripción generada automáticamente

**Cambio de nombre de las variables**

Debido al idioma de origen y la complejidad para entender el contenido de la variable, utilizamos la función rename para realizar el cambio de nombre en las columnas, quedando de la siguiente manera la equivalencia:



**Aplicamos la funcion str para ver los detalles y caracteristicas de los datos del dataset.**



Vemos nuevamente los datos pertenecientes a este dataset, pero ya con nuestro cambio en el nombre de variables, cuyos nombres son mucho más intuitivos.

Texto, Carta

Descripción generada automáticamente

**Mediante la siguiente transformacion convertiremos nuestras variabler chr a factor**

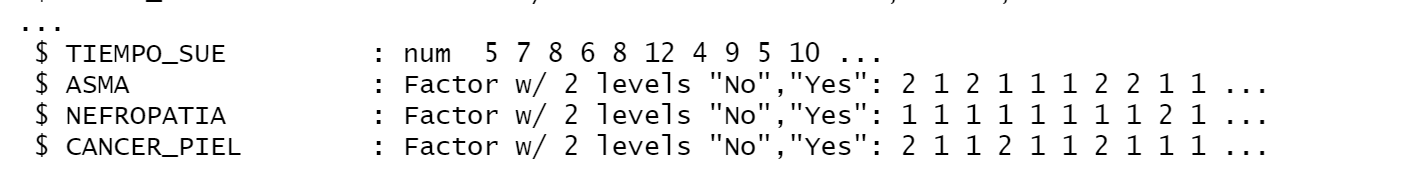
Texto

Descripción generada automáticamente

Podemos ver como nuestras variables se vuelven categricas y por cada variable sus niveles.

Texto

Descripción generada automáticamente



**Verificando datos nulos y na**

Debido a la importancia de identificar y trabajar los datos nulos o perdidos, crearemos nuevos dataset que mediante el operador lógico TRUE realizarán un filtro, esto con el fin de detectarlos y que no afecten nuestras predicciones.

Interfaz de usuario gráfica, Aplicación

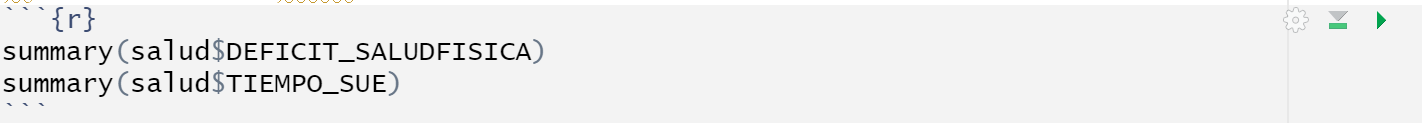
Descripción generada automáticamente

Vemos que las filas que cumplen la condicion son 0, por lo tanto descartamos que nuestro dataset tenga na, nan o infinite.

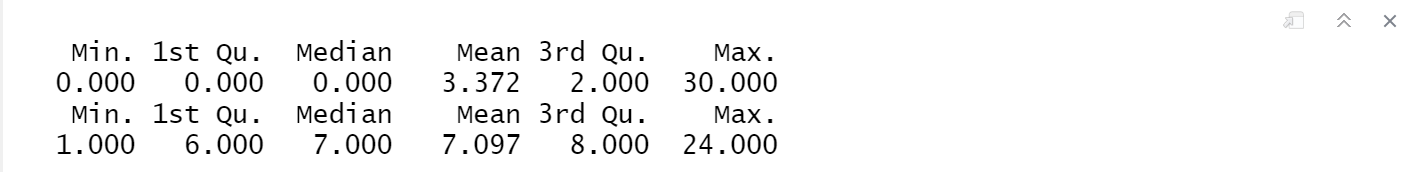
Tabla

Descripción generada automáticamente

Aplicaremos el comando summary para ver un resumen de nuestras dos variables escogidas (DEFICIT\_SALUDFISICA y TIEMPO\_SUE), este nos entrega el minimo,maximo, el primer y tercer quintil, la media y mediana. Si existieran datos “na” este comando arrojaráa la cantidad, como no figura se desprende que no existen.



El resumen de datos acontinuacion:



Es esencial saber cual es el dato que más se repite para ver las tendencias y lograr realizar un analísis en mayor profundidad, para esto creamos una función llamada mode\_deficitsaludfis y mode\_tiemposueno.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

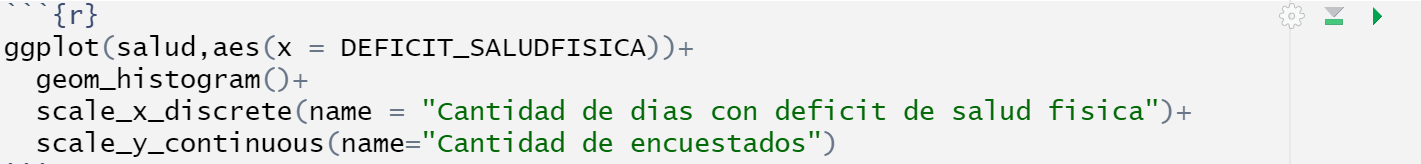
Descripción generada automáticamente

Podemos ver que el dato que más se repite es 7 en la cantidad de horas diarias que duerme la persona, en lo que respecta a la cantidad de días que la persona no rinde físicamente o no tienen una buena salud física es cero



Podríamos desprender de esto, que si bien existe una gran cantidad de datos de personas que indican que no  tienen días que no puedan realizar actividades, es muy importante estudiar los restantes a fin de determinar que les provoca este trastorno.

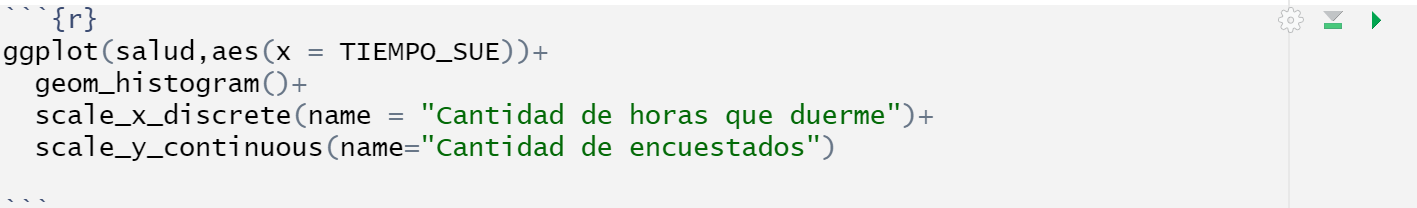
Para ver esta información de mejor manera utilizaremos la función ggplot  en dichas variables y poder tener una visualización de ellas.

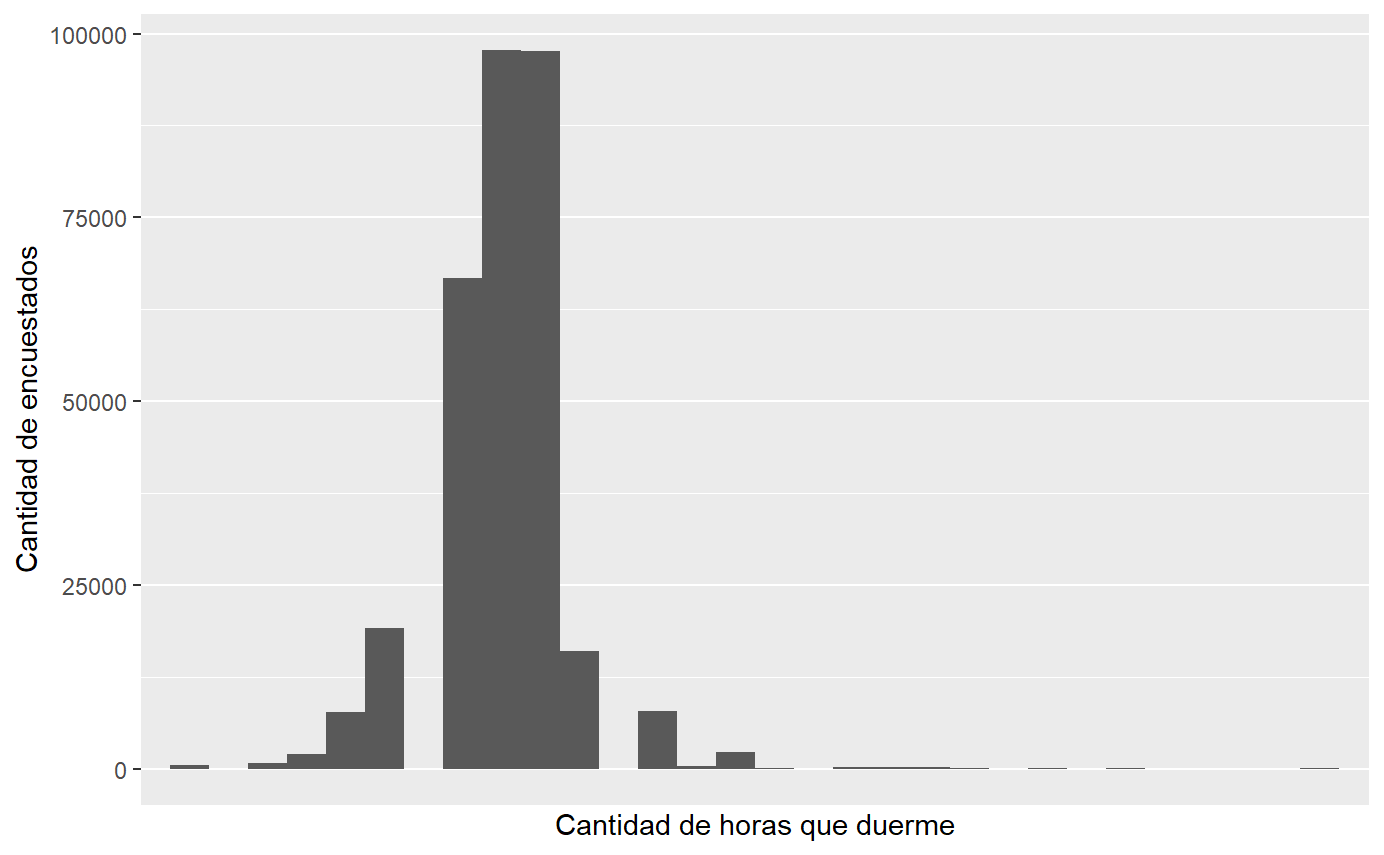


Gráfico

Descripción generada automáticamente

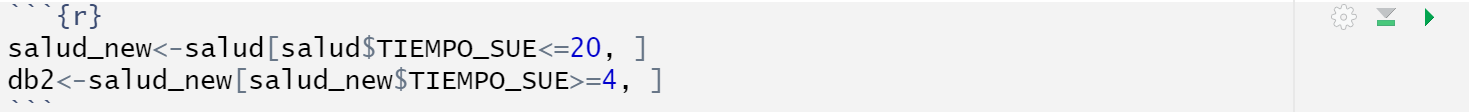
Aplicaremos lo mismo para nuestra varible de horas de sueño





Realizaremos una reestructuración de los datos, ya que se considera que una persona no podría dormir las 24 horas, por ende  permitiremos valores de hasta 18 horas, y sacaremos los datos de aquellos que duermen solo una hora, ya que no es sano mantenerse de aquella manera.

En esta nueva creación de dataset, salud\_new representará el conjunto de datos con los valores donde el tiempo de sueño sea menor o igual a 20 horas, luego de realizar este filtro



En nuestro primer filtro vemos que se redujo a 319751 filas o datos ya que eliminó todos aquellos donde el tiempo de sueño sea menor o igual a 20.



En un segundo filtro creamos un dataset llamado db2, al cual se le aplica el filtro con tiempo de sueño mayor o igual a cuatro. Con ambos filtros aplicados nos quedamos con una totalidad de 316420 datos.



**Correlación de los datos**

Dado que queremos determinar el coeficiente de correlación entre las dos variables mencionadas anteriormente, utilizaremos la función cor. En el eje y consideraremos la variable DEFICIT\_SALUDFISICA, ya que es la variable dependiente y en el eje x tendremos la variable TIEMPO\_SUE.



La correlación entre estas variables es extremadamente baja muy cercana al cero, eso significa que los datos se encuentran muy dispersos y en cuanto al signo al ser negativa significa que se relacionan en sentido inverso.

Texto

Descripción generada automáticamente

Ahora realizamos la gráfica con el dataset db2 para poder tener mayor claridad de cómo se encuentran los datos. Graficamos con este dataset, ya que con ella trabajaremos hasta el final de las regresiones.

Texto

Descripción generada automáticamente

Imagen que contiene Gráfico

Descripción generada automáticamente

**Modelos de regresión**

Con el fin de describir la relacion entre las variables utilizaremos el analisis de regresion. En el primer caso trabajaremos con dos variables, la predictora y la de respuesta. En los siguientes regresiones analizaremos casos los que involucran una mayor cantidad de variables.

Para interpretar el valor p-value nos valdremos del valor 0.05. Este valor evalua la hipotesis nula, si este valor es <=0.05 significa que podemos rechazar la hipotesis nula, es decir el valor de esta categoria definira si la variable es significativa y entrega valor significativo al modelo. Por el contrario si el valor es 0.05> sugiere que no afecta ni es significante para este modelo de regresion.

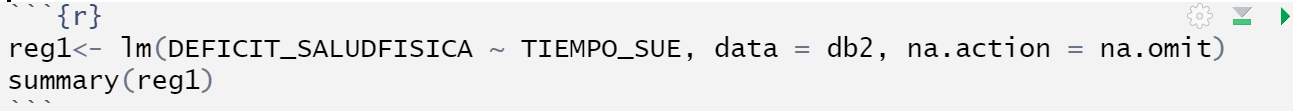
**Regresión 1**

En la regresión 1 (reg1) se utiliza como variable predictora el tiempo de sueño de una persona y como variable de respuesta la cantidad de días que la persona no rindió físicamente.

El resultado nos indica un valor p-value de 2e-16 equivalente a 0.000015258 un valor por debajo de 0.05 lo que significa que la variable TIEMPO\_SUE si es significante para nuestro modelo.

El coeficiente por otro lado nos indica que si aumenta la cantidad de horas que duerme la persona, disminuirá en 0.22965 los días que presente una déficit en su salud física.

Según el cálculo del R cuadrado, es de 0.0015, lo que significa que el modelo explica de una manera insignificante a la variable dependiente.



Texto, Carta

Descripción generada automáticamente

**Regresión 2**

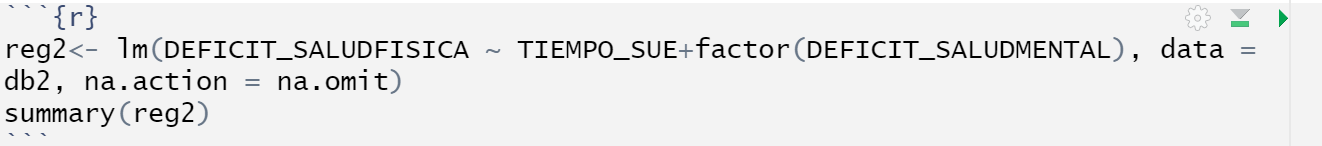
La elección de las variables participantes de esta regresión se basan en la fuerte idea de qué presentar problemas de salud mental influye en desarrollar enfermedades o una disminución de la salud física. Existen estudios que afirman que la depresión aumenta el riesgo de sufrir una enfermedad cardiovascular y cáncer. Podemos asociar los siguientes síntomas físicos con la depresión:

Depresión -> Cansado/Agotado

Depresión -> Cambio en el apetito (Obesidad)

Depresión -> Dolor

Depresión -> Cambio de comportamiento( Haciendo menos cosas)



Para esta regresión se agregó a la regresión anterior el factor DEFICIT\_SALUDMENTAL, esto con la idea de ver si existe relación entre los días que la persona no presenta una buena salud mental con los días que no presenta buena salud física.

El coeficiente por otro lado al tratarse de un factor entrega los coeficientes de acuerdo a la cantidad de días, los cuales son 30 distintos. Solo el valor 1 toma un valor mayor a 0.05 por lo tanto no es significativo el resto presenta valores menores por lo tanto son significativos en nuestro modelo.

Según el cálculo del R cuadrado es de 0.075, aumenta el porcentaje de fidelidad del modelo, sin embargo un 7,93% está muy por debajo del 65% mínimo.

Texto

Descripción generada automáticamente

Tabla

Descripción generada automáticamente

Texto

Descripción generada automáticamente

**Regresión 3**

La idea de la tercera regresión es reunir todos los factores que indican enfermedades, ya sea las que padecen al momento de ser consultados o las que generaron secuelas luego de haberlas padecido.

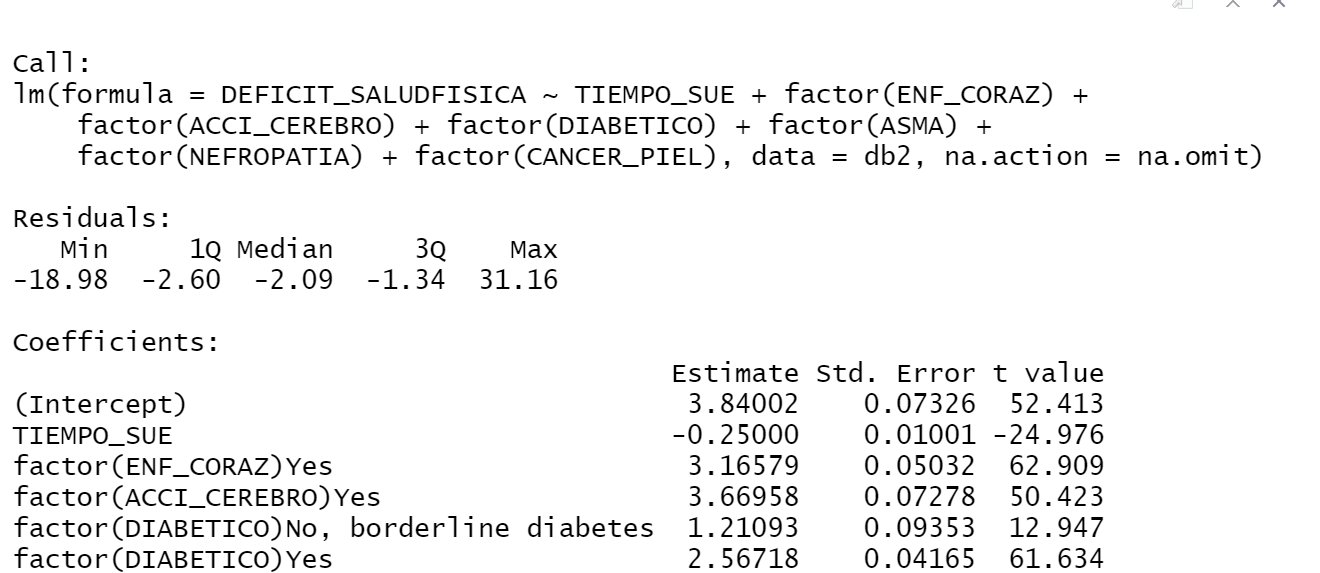
**Texto

Descripción generada automáticamente**

Los valores de p-value en esta regresión indican que todas las variables a excepción del factor DIABETICO yes (during pregnancy) son significantes en este modelo por tener un valor menor o igual a 0.05

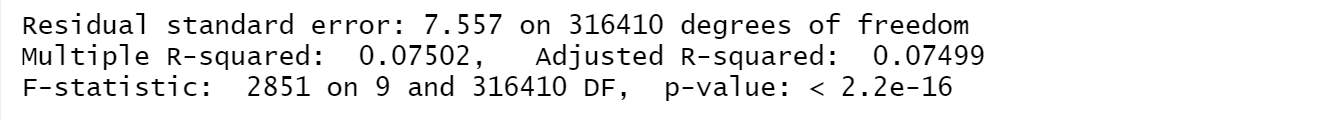
Analizaremos los coeficientes de las variables que no fueron detalladas anteriormente. En lo que respecta a las enfermedades al corazón tienen si la respuesta YES aumenta en 3.16579 los dias que no tendrá una buena actividad física, en cerebrovascular aumenta en 3.66958, si se encuentra al borde de la diabetes aumenta en 1.21093 dias, si es diabetico aumentara 2.56718, si la respuesta es al borde de la diabetes 1.21093, en el caso del asma aumenta 2.22604 días, nefropatía aumenta 3.99831 y finalmente si tiene cáncer de piel aumenta 0.51012

En esta regresión el valor de R cuadrado es de 0.07502, inferior a la regresión anterior y muy baja.

****

**Una captura de pantalla de un celular

Descripción generada automáticamente**

****

**Regresión 4**

En esta nueva regresión adicionamos nuevos factores incluyendo la variable DIFFCAMINAR, ya que analizaremos si la dificultad para caminar es el principal factor que no permite al individuo tener una salud física y afecta su rutina.

**Texto, Carta

Descripción generada automáticamente con confianza media**

Debido a que ya comentamos cómo afecta el factor salud mental, mencionaremos cómo influye la dificultad para cambiar, esto quiere decir que si la respuesta es afirmativa aumentará en 8.925033 días.

El valor de p-value en la variable DEFFCAMINAR es menor a 0.05 así que se considera significativo para nuestro modelo.

El R cuadrado aumentó en comparación con las anteriores regresiones llegando a un 0.2279 equivalente a 22.79 % , valor poco robusto pero mejor a los anteriores.

**Una captura de pantalla de un celular

Descripción generada automáticamente**

**Un periódico con texto

Descripción generada automáticamente con confianza media**

**Texto

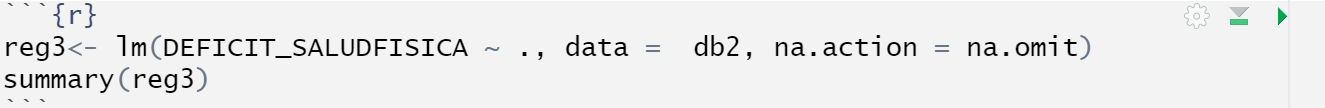
Descripción generada automáticamente**

**Regresion 5**

Finalmente nuestra ultima regresion contiene como variables independientes todas las contenidas en el dataset a excepcion de la variable dependiente. La idea es verificar si todas las variables de manera conjunta logran tener la mejor precision y limpiar los datos.

Podemos decir de la variable sexo que tiene un valor mayor a 0.05 por lo que no es relevante en nuesto modelo y podriamos precindir de el, al igual que el factor raza de otro tipo, la categoria de edad de 25 a 29, raza blanca y diabeticos during pregnancy.

El R cuadrado es 0.3968 equivalente a 39.68% el mayor valor que pudimos obtener de nuestras regresiones, sin embargo no logramos un valor lo suficientemente alto, este valor no se considera bueno para modelar los datos.

****

**Texto

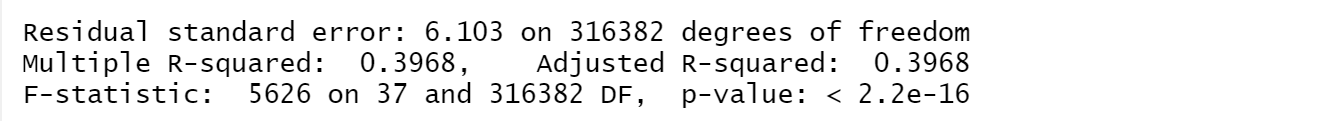
Descripción generada automáticamente**

**Tabla

Descripción generada automáticamente**

**Tabla

Descripción generada automáticamente**

****

### Conclusión

Gran parte importante de este desarrollo era lograr encontrar nuestro modelo que presente mejor porcentaje de la variación en la variable dependiente, esto lo logramos con R cuadrado.

En lo que respecta a nuestras hipótesis finalmente rechazamos la hipótesis nula, lo que quiere decir que si efectivamente existe una relación entre la cantidad de horas que duerme una persona con los días que no puede tener una buena salud física. Esto sí, es importante mencionar que esta variable debe ir acompañada de otros factores para que logre explicar de mejor manera la cantidad de días. Por lo que nos quedamos con la última regresión que incluye todas las variables tanto numéricas como factor.

Podemos concluir entonces que estamos tratando de predecir un comportamiento humano , por lo tanto el R-cuadrado podría no llegar siempre a tener siempre un porcentaje alto. En el caso de nuestro modelo final si bien no es alto tiene predictores estadísticamente significativos, lo que permite tener asociaciones importantes entre las variables independientes y la dependiente.