2021

ACTIVIDAD 1

ANÁLISIS ESTADÍSTICO COMPLETO CON RSTUDIO

CRISTINA GONZÁLEZ CELADA

UNIVERSIDAD INTERNACIONAL DE VALENCIA
MÁSTER UNIVERSITARIO EN BIG DATA Y DATA SCIENCE
Estadística Avanzada

CONTENIDOS

1. OBJETIVOS

El objetivo principal de esta práctica consiste en la realización de un análisis estadístico completo de un dataset, o base de datos, programando con la herramienta RStudio. Este análisis se realizará completando un análisis descriptivo univariante y bivariante de la base de datos, además de estimando y validando un modelo de regresión. Se interpretarán los resultados obtenidos y se analizará, también, el potencial del modelo, junto con sus limitaciones y posibles mejoras.

Este análisis estadístico se lleva a cabo siguiendo el siguiente esquema:

1. Se explicará de forma detallada el **contexto** del dataset escogido, además de los posibles objetivos específicos que se pretenden alcanzar.
2. Se analizará la **base de datos** elegida, comentando las dimensiones de ésta, los tipos de datos que contiene, etc.
3. Se realizará un **análisis descriptivo**, donde se reflejen parámetros como la media, moda, varianza, desviación típica de los datos, entre otros.
4. Se estimará y validará un **modelo de regresión**, que permita realizar predicciones sobre los datos.
5. Se explicará con detalle las **conclusiones** alcanzadas tras el análisis.
6. Finalmente, se expondrán, además, los posibles **trabajos futuros** del dataset, con sus respectivas **limitaciones** y **mejoras** que podrían aplicarse.

2. CONTEXTO

A lo largo del tiempo la sociedad se ha visto cada vez más envuelta en el bienestar de los propios individuos que la componen. Muchas personas se preocupan en hacer más ejercicio, comer sano, tener una vida social satisfactoria, alcanzar metas personales y/o profesionales, mantener una estabilidad emocional… En definitiva, alcanzar y mantener unos buenos hábitos. Son tareas que requieren esfuerzo y sacrificio, por lo que es un objetivo no apto para todas las personas que se lo proponen, ya que es determinante tener fuerza de voluntad y constancia, entre otras características.

Es en este ámbito en el que el presente trabajo se va a focalizar, observando y analizando un dataset de individuos que han contestado determinadas preguntas, de una encuesta anónima, relacionadas con su forma de vida. Esta forma de vida viene definida por las siguientes cinco dimensiones, explicadas a continuación:

- Cuidado del cuerpo
- Cuidado de la mente
- Experiencias personales
- Conexión con el mundo
- Significado de la vida

El *cuidado del cuerpo* refleja la forma física y la periodicidad de práctica de deporte realizada por el individuo, mientras que el *cuidado de la mente* trata de representar cómo acoge las emociones positivas que se pueden presentar en los diferentes ámbitos de su vida. Las *experiencias personales* intentan medir su habilidad o capacidad de alcanzar logros y hacer crecer su experiencia personal. Esto puede estar directamente relacionado también con su *conexión con el mundo*, ya que vive en sociedad al estar conectado a través de los lazos sociales. El *significado de la vida* que le da cada persona puede ser evaluado a través de su grado de compasión, generosidad y forma de vivir su vida como un sueño.

A partir de estos datos, enmarcados en esas cinco dimensiones, se pretende comprobar la correlación existente entre las diferentes variables que intervienen en la encuesta y, finalmente, si esa puntuación obtenida está basada en un modelo válido.

3. BASE DE DATOS

La comunidad online de científicos de datos y profesionales del aprendizaje automático, Kaggle, ofrece diversos datasets, en función de la temática, usabilidad, autor, tareas a realizar, etc.

En este caso particular, se ha obtenido de esta comunidad un dataset que contiene datos recogidos de una encuesta anónima, cuya temática trata sobre los buenos hábitos y los estilos de vida de las personas encuestadas. El usuario contesta 22 preguntas, entre las cuales puede indicar su correo electrónico y dar su consentimiento para el tratamiento de los datos y, en base a sus respuestas, se genera una puntuación total que determina cuán buenos son los hábitos de esa persona, hábitos que quedan enmarcados en las 5 dimensiones mencionadas en el apartado anterior.

3.1 CARACTERÍSTICAS PRINCIPALES

Esta base de datos consta de **15.972 observaciones** (filas) y **24 variables** (columnas). Las posibles variables sobre las que se va a poder realizar el análisis estadístico recogen la siguiente información:

- TIMESTAMP
  - Fecha y hora de la realización de la encuesta.
  - Los valores máximo y mínimo de esta variable son 01/08/2015 y 14/03/2021, respectivamente, por lo que el dataset contiene datos de 6 años diferentes.
- FRUITS_VEGGIES
  - Número de piezas de fruta y/o verdura ingeridas al día.
  - Es una variable cuantitativa que puede tomar valores de entre 0 y 5 piezas al día.
- DAILY_STRESS
  - Cantidad de estrés diario sufrido.
  - Es una variable categórica que puede tomar valores comprendidos entre el 0 y el 5, correspondiéndose estos valores a la siguiente escala o equivalencia:
    - 0 = No mucho estrés
    - 5 = Mucho estrés

    Dado que los valores 1, 2, 3 y 4 no reflejan el grado de estrés sufrido, se puede 'interpolar' el significado de la siguiente forma:
    - 1 = Estrés muy leve
    - 2 = Estrés leve
    - 3 = Estrés moderado
    - 4 = Estrés elevado

- PLACES_VISITED
    - Lugares nuevos visitados al año.
    - Es una variable cuantitativa que puede tomar valores de entre 0 y 10 lugares visitados al año.
- CORE_CIRCLE
    - Número de personas cercanas, referido al círculo cercano de una persona (familiares, amigos, compañeros de trabajo, etc.)
    - Es una variable cuantitativa que puede tomar valores de entre 0 y 10 personas cercanas.
- SUPPORTING_OTHERS
    - Número de personas a las que se presta algún tipo de ayuda (económica, social, moral, etc.) con el objetivo de ayudarles a conseguir una mejor vida, siendo mentor, entrenador, desarrollador, promotor altruista, etc.
    - Es una variable cuantitativa que puede tomar valores de entre 0 y 10 personas.
- SOCIAL_NETWORK
    - Número de personas al día con las que se interacciona.
    - Es una variable cuantitativa que puede tomar valores de entre 0 y 10 personas al día.
- ACHIEVEMENT
    - Logros anuales memorables (tanto personales como profesionales)
    - Es una variable cuantitativa que puede tomar valores de entre 0 y 10 logros al año.
- DONATION
    - Número de veces al año que se presta ayuda desinteresadamente (donando tiempo o dinero a causas benéficas, por ejemplo)
    - Es una variable cuantitativa que puede tomar valores de entre 0 y 5 veces al año.
- BMI_RANGE
    - Índice de masa corporal (en ingles: Body Mass Index)
    - Es una variable categórica que puede tomar los siguientes dos valores:
        - 1 = BMI por debajo de 25
        - 2 = BMI por encima de 25
- TODO_COMPLETED
    - Evaluación de satisfacción en cuanto a cumplimiento de objetivos en una lista de tareas semanales.

  - Es una variable categórica que puede tomar valores comprendidos entre el 0 y el 10, correspondiéndose estos valores a la siguiente escala o equivalencia:
    - 0 = No muy bien
    - 10 = Muy bien

- FLOW
  - Horas diarias en las que se experimenta un estado de 'flow', es decir, en las que el estado mental de una persona es de inmersión total en una actividad concreta.
  - Es una variable cuantitativa que puede tomar valores de entre 0 y 10 horas al día.
- DAILY_STEPS
  - Número de pasos diarios caminados, medido en miles de pasos al día.
  - Es una variable cuantitativa que puede tomar valores de entre 1 y 10 miles de pasos al día.
- LIVE_VISION
  - Número de años en los que una persona mantiene la misma visión de la vida.
  - Es una variable cuantitativa que puede tomar valores de entre 0 y 10 años.
- SLEEP_HOURS
  - Número de horas de sueño dormidas por una persona.
  - Es una variable cuantitativa que puede tomar valores de entre 1 y 10 horas al día.
- LOST_VACATION
  - Número de días perdidos al año de vacaciones, bien por no usar esos días, por ser usados en otro año, o por haber sido interrumpidas por estrés, trabajo, etc.
  - Es una variable cuantitativa que puede tomar valores de entre 0 y 10 días perdidos al año.
- DAILY_SHOUTING
  - Número de veces por semana que una persona grita o experimenta un estado de apatía, abatimiento, disgusto o tristeza.
  - Es una variable cuantitativa que puede tomar valores de entre 0 y 10 veces por semana.
- SUFFICIENT_INCOME
  - Grado de satisfacción de una persona en cuanto a los ingresos recibidos para poder cubrir las necesidades básicas.
  - Es una variable categórica que puede tomar los siguientes dos valores:
    - 1 = No lo suficiente
    - 2 = Suficiente

- PERSONAL_AWARDS
    - Número de reconocimientos totales que recibe una persona a lo largo de su vida, bien personales, bien profesionales, en forma de diplomas, grados, certificados, acreditaciones, premios, publicaciones, presentaciones, medallas, etc.
    - Es una variable cuantitativa que puede tomar valores de entre 0 y 10 reconocimientos.
- TIME_FOR_PASSION
    - Horas diarias que una persona invierte en una actividad que le apasiona, enfocado en la contribución a una mejor causa: salud, educación, paz, desarrollo social, etc.
    - Es una variable cuantitativa que puede tomar valores de entre 0 y 10 horas al día.
- WEEKLY_MEDITATION
    - Número de veces por semana que una persona invierte en pensar en sí mismo (meditando, rezando, actividades de relajación, etc.)
    - Es una variable cuantitativa que puede tomar valores de entre 0 y 10 veces por semana.
- AGE
    - Rango de edad del individuo que realiza la encuesta. Existen 4 rangos de edad:
        - Menores a 20 años (less than 20)
        - Entre 21 y 35 años (21 to 35)
        - Entre 36 y 50 años (36 to 50)
        - Mayores a 51 años (51 or more)
    - Es una variable categórica que puede tomar los valores mencionados anteriormente.
- GENDER
    - Género de la persona que realiza la encuesta. Dos tipos de género:
        - Femenino (Female)
        - Masculino (Male)
    - Es una variable categórica que puede tomar los valores mencionados anteriormente.
- WORK_LIFE_BALANCE_SCORE
    - Puntuación total obtenida según el resto de apartados analizados.
    - Es una variable cuantitativa que queda acotada entre 0 y 1000.

El dataset se obtiene de una forma bastante 'limpia', es decir, no se observa falta de información sobre ninguna observación (missing values o valores nulos), aunque es cierto que necesita del tratamiento de algún campo como Daily_Stress, puesto que analizando previamente los datos, se comprueba que existe una observación con un valor erróneo (si éste debe ser un valor comprendido entre 0 y 5, el valor erróneo presenta un dato de tipo fecha)

Además, no existe un campo que indique que cada registro es información de un único usuario, por lo que la muestra real de la encuesta es ambigua. Cada variable tiene una unidad de medida diferente: horas, días/año, número de personas, etc. Al disponer de acceso a la propia encuesta se ha observado que las escalas de puntuación de cada pregunta son diferentes: 0-5, 0-10, 1-2,... Por lo que el dataset al final muestra este tipo de datos en los resultados: números enteros, excepto la puntuación final obtenida, que es de tipo flotante.

En resumen, se podría decir que la base de datos elegida consta de las siguientes características:

Nº OBSERVACIONES	15.972
Nº VARIABLES	24

DIMENSIONES*	Descripción
Cuerpo	Cuerpo sano: forma física y hábitos saludables
Mente	Mente sana: emociones positivas
Exp. Personal	Capacidad de hacer crecer su experiencia y logros
Conexión Mundo	Fuerza de red social e inclinación por descubrimiento del mundo
Significado Vida	Compasión, generosidad, vivir la vida como en el sueño

*Algunas variables pueden admitir varias dimensiones.

TIPOLOGÍAS	Descripción
1	Cuantitativa
2	Categórica

RESUMEN	Descripción
Ud. Temporales	Horas/día; horas/noche; días/año; años; etc.
Ud. Cantidad	Personas; etc.
Puntuaciones	Valoración en rangos de puntuación: 1-2, 0-5, 0-10, etc.

NUM.	VARIABLE	TRADUCCIÓN	UNIDADES	TIPOLOGÍA	DIMENSIÓN
1	Timestamp	Fecha	DD/MM/AAAA; DD/MM/AAAA HH:MM	-	-
2	FRUITS_VEGGIES	Frutas_Verduras	Piezas/día (0-5)	1	Cuerpo
3	DAILY_STRESS	Estres_Diario	Puntuación (0-5)	2	Mente
4	PLACES_VISITED	Lugares_Visitados	Lugares/año (0-10)	1	Exp. Personal
5	CORE_CIRCLE	Circulo_Personal	Personas (0-10)	1	Conexión Mundo
6	SUPPORTING_OTHERS	Ayudar_Otros	Personas (0-10)	1	Significado Vida
7	SOCIAL_NETWORK	Redes_Sociales	Personas/día (0-10)	1	Conexión Mundo
8	ACHIEVEMENT	Retos_Memorables	Logros/año (0-10)	1	Exp. Personal
9	DONATION	Donacion_Causas	Veces/año (0-5)	1	Significado Vida
10	BMI_RANGE	Indice_Masa_Corporal	Puntuación (1-2)	2	Cuerpo
11	TODO_COMPLETED	Lista_Tareas	Puntuación (0-10)	2	Exp. Personal
12	FLOW	Estado_Inmersion	Horas/día (0-10)	1	Conexión Mundo
13	DAILY_STEPS	Pasos_Diarios	Miles de pasos/día (0-10)	1	Cuerpo
14	LIVE_VISION	Vision_Vida	Años (0-10)	1	Significado Vida
15	SLEEP_HOURS	Horas_Sueño	Horas/noche (1-10)	1	Cuerpo
16	LOST_VACATION	Vacaciones_Perdidas	Días/año (0-10)	1	Exp. Personal
17	DAILY_SHOUTING	Dias_Gritando	Veces/semana (0-10)	1	Conexión Mundo
18	SUFFICIENT_INCOME	Ingresos_Basicos	Puntuación (1-2)	2	Exp. Personal
19	PERSONAL_AWARDS	Reconocimientos_Personales	Reconocimientos/vida (0-10)	1	Exp. Personal
20	TIME_FOR_PASSION	Tiempo_Pasiones	Horas/día (0-10)	1	Significado Vida
21	WEEKLY_MEDITATION	Meditacion_Semanal	Veces/semana (0-10)	1	Mente
22	AGE	Edad	Rangos	2	-
23	GENDER	Genero	M-F	2	-
24	WORK_LIFE_BALANCE_SCORE	Puntuación	-	-	-

3.2 DATASET EMPLEADO EN EL ANÁLISIS

Dado que el dataset original contempla una gran cantidad de datos se va a proceder a realizar un filtrado de la información, con el objetivo de acotar la muestra de datos. En este sentido, se va a focalizar el estudio en los datos del año 2016. Esto hace que el número de variables sea el mismo que la base de datos original (**24 variables**), mientras que el número de observaciones disminuya de 15.972 a **3.328 observaciones**. Como se ha explicado en el contexto de esta memoria, se va a analizar la validez de la puntuación total obtenida a partir del resto de variables (sin incluir la variable Timestamp, ya que todas las fechas de estos datos se corresponderán con el año 2016 y no es una variable relevante para el estudio).

El proceso de filtrado de datos se ha llevado a cabo mediante la herramienta Microsoft SQL Server y los editores de texto Microsoft Excel y Bloc de Notas, a partir del siguiente procedimiento:

1. Importación del dataset original en una única tabla, dentro de Mmicrosoft SQL Server, denominada dbo.Wellbeing_and_lifestyle_data_Kaggle.

2. Almacenado de los datos relativos al año 2016 en una nueva tabla, denominada dbo.Datos2016. Estos datos conservan la misma tipología que los datos originales. Se lleva a cabo mediante la siguiente sentencia SQL:

```
IF OBJECT_ID('dbo.Datos2016') IS NOT NULL DROP TABLE dbo.Datos2016
SELECT
	*
INTO dbo.Datos2016
FROM
	dbo.Wellbeing_and_lifestyle_data_Kaggle
WHERE
	Timestamp like '%/16'
```

3. Exportación de todos los datos de la tabla anterior a una hoja de cálculo Excel.
4. Exportación a fichero de formato txt. Este último fichero, denominado 'Dataset_Datos_2016.txt', será el que se utilice en Rstudio.

4. ANÁLISIS DESCRIPTIVO

En primer lugar, se tiene que establecer una variable en torno a la que gire todo el análisis descriptivo. Por ello, el primer paso que se lleva a cabo es la realización de una matriz de correlación entre todas las variables numéricas del dataset. Esta matriz permite observar, de forma general, las relaciones (directas o inversamente proporcionales) entre las diferentes variables del dataset.

El procedimiento en Rstudio es el siguiente:

1. Instalación y carga de las librerías utilizadas.

```
#IMPORTACIÓN DE LIBRERÍAS NECESARIAS:
  #Instalación de paquetes:
install.packages('dplyr')       #Enlace librería: https://swcarpentry.github.io/r-novice-gapminder-es/13-dplyr/
install.packages("corrplot")    #Enlace librería: https://rpubs.com/camilamila/correlaciones
install.packages("ISLR")
install.packages("car")
install.packages("ggplot2")

  #Carga de paquetes:
library(dplyr)
library(corrplot)
library(ISLR)
library(car)
library(ggplot2)
```

2. Apertura del dataset (datos del año 2016) y almacenamiento del mismo en la variable 'db'. Además, se almacena el número de observaciones totales en la variable 'muestra':

```
#APERTURA DEL DATASET:
getwd()
setwd("E:/VIU/Asignaturas/6_Estadística Avanzada/Actividades/Actividad_1/Entrega")

  #Dataset completo (año 2016):
db <- read.table("Dataset_Datos_2016.txt", header = TRUE, sep = ",", dec = ".", skip = 0)
  #Número de observaciones:
muestra <- dim(db)[1]
```

3. Definición de un dataframe con las observaciones y variables numéricas (denominado 'dbn'). En este caso concreto se excluyen las variables Timestamp, AGE y GENDER. El resto de variables categóricas, como DAILY_STRESS, BMI_RANGE, TODO_COMPLETED y SUFFICIENT_INCOME, se pueden incluir porque su valor es de tipo numérico, aunque se deberá tener especial cuidado en el tratamiento de las mismas en la posterior estimación del modelo de regresión. Además, se genera un dataframe por cada variable en torno a las que girará el estudio descriptivo: variables WORK_LIFE_BALANCE_SCORE (wlbs) y ACHIEVEMENT (achiev).

```
#DEFINICIÓN DE SUBCONJUNTOS DE VARIABLES A UTILIZAR:
  #Dataset sin: Timestamp.
db2 <- select(db,FRUITS_VEGGIES,DAILY_STRESS,PLACES_VISITED,CORE_CIRCLE,SUPPORTING_OTHERS,SOCIAL_NETWORK,ACHIEVEMENT
              ,DONATION,BMI_RANGE,TODO_COMPLETED,FLOW,DAILY_STEPS,LIVE_VISION,SLEEP_HOURS,LOST_VACATION,DAILY_SHOUTING
              ,SUFFICIENT_INCOME,PERSONAL_AWARDS,TIME_FOR_PASSION,WEEKLY_MEDITATION,AGE,GENDER,WORK_LIFE_BALANCE_SCORE)

  #Dataset con datos numéricos: sin AGE, GENDER.
dbn <- select(db2,FRUITS_VEGGIES,DAILY_STRESS,PLACES_VISITED,CORE_CIRCLE,SUPPORTING_OTHERS,SOCIAL_NETWORK,ACHIEVEMENT
              ,DONATION,BMI_RANGE,TODO_COMPLETED,FLOW,DAILY_STEPS,LIVE_VISION,SLEEP_HOURS,LOST_VACATION,DAILY_SHOUTING
              ,SUFFICIENT_INCOME,PERSONAL_AWARDS,TIME_FOR_PASSION,WEEKLY_MEDITATION,WORK_LIFE_BALANCE_SCORE)

  #Dataset con datos de WORK_LIFE_BALANCE_SCORE y ACHIEVEMENT:
wlbs <- select(db2,WORK_LIFE_BALANCE_SCORE)
achiev <- select(db2,ACHIEVEMENT)
```

4. Obtención de la matriz de correlación, con una precisión o redondeo de 2 decimales.

	FRUITS_VEGGIES	DAILY_STRESS	PLACES_VISITED	CORE_CIRCLE	SUPPORTING_OTHERS	SOCIAL_NETWORK	ACHIEVEMENT	DONATION	FLOW	DAILY_STEPS	LIVE_VISION
FRUITS_VEGGIES	1.00	-0.07	0.25	0.14	0.18	0.10	0.15	0.19	0.10	0.23	0.08
DAILY_STRESS	-0.07	1.00	-0.12	-0.13	-0.03	0.00	-0.13	-0.05	-0.14	-0.03	-0.13
PLACES_VISITED	0.25	-0.12	1.00	0.23	0.25	0.13	0.25	0.19	0.14	0.19	0.13
CORE_CIRCLE	0.14	-0.13	0.23	1.00	0.34	0.31	0.29	0.22	0.25	0.14	0.20
SUPPORTING_OTHERS	0.18	-0.03	0.25	0.34	1.00	0.33	0.37	0.37	0.27	0.15	0.23
SOCIAL_NETWORK	0.10	0.00	0.13	0.31	0.33	1.00	0.26	0.15	0.24	0.22	0.16
ACHIEVEMENT	0.15	-0.13	0.25	0.29	0.37	0.26	1.00	0.24	0.40	0.23	0.30
DONATION	0.19	-0.05	0.19	0.22	0.37	0.15	0.24	1.00	0.19	0.10	0.16
FLOW	0.10	-0.14	0.14	0.25	0.27	0.24	0.40	0.19	1.00	0.17	0.30
DAILY_STEPS	0.23	-0.03	0.19	0.14	0.15	0.22	0.23	0.10	0.17	1.00	0.11
LIVE_VISION	0.08	-0.13	0.13	0.20	0.23	0.16	0.30	0.16	0.30	0.11	1.00
SLEEP_HOURS	0.10	-0.12	0.14	0.05	0.02	-0.05	0.04	0.03	0.02	0.02	0.03
LOST_VACATION	-0.06	0.21	-0.13	-0.06	0.00	0.08	-0.01	-0.01	0.01	-0.03	-0.02
DAILY_SHOUTING	-0.04	0.35	-0.08	-0.10	-0.05	0.00	-0.08	-0.09	-0.08	0.00	-0.06
PERSONAL_AWARDS	0.15	-0.06	0.30	0.26	0.34	0.21	0.38	0.26	0.22	0.19	0.20
TIME_FOR_PASSION	0.15	-0.15	0.16	0.22	0.33	0.20	0.37	0.20	0.48	0.15	0.27
WEEKLY_MEDITATION	0.21	-0.17	0.20	0.08	0.09	-0.03	0.14	0.13	0.12	0.18	0.12
WORK_LIFE_BALANCE_SCORE	0.43	-0.37	0.53	0.50	0.55	0.39	0.58	0.46	0.49	0.42	0.45

	SLEEP_HOURS	LOST_VACATION	DAILY_SHOUTING	PERSONAL_AWARDS	TIME_FOR_PASSION	WEEKLY_MEDITATION	WORK_LIFE_BALANCE_SCORE
FRUITS_VEGGIES	0.10	-0.06	-0.04	0.15	0.15	0.21	0.43
DAILY_STRESS	-0.12	0.21	0.35	-0.06	-0.15	-0.17	-0.37
PLACES_VISITED	0.14	-0.13	-0.08	0.30	0.16	0.20	0.53
CORE_CIRCLE	0.05	-0.06	-0.10	0.26	0.22	0.08	0.50
SUPPORTING_OTHERS	0.02	0.00	-0.05	0.34	0.33	0.09	0.55
SOCIAL_NETWORK	-0.05	0.08	0.00	0.21	0.20	-0.03	0.39
ACHIEVEMENT	0.04	-0.01	-0.08	0.38	0.37	0.14	0.58
DONATION	0.03	-0.01	-0.09	0.26	0.20	0.13	0.46
FLOW	0.02	0.01	-0.08	0.22	0.48	0.12	0.49
DAILY_STEPS	0.02	-0.03	0.00	0.19	0.15	0.18	0.42
LIVE_VISION	0.03	-0.02	-0.06	0.20	0.27	0.12	0.45
SLEEP_HOURS	1.00	-0.10	-0.09	0.05	0.04	0.17	0.20
LOST_VACATION	-0.10	1.00	0.07	-0.05	-0.04	-0.13	-0.24
DAILY_SHOUTING	-0.09	0.07	1.00	-0.03	-0.10	-0.08	-0.28
PERSONAL_AWARDS	0.05	-0.05	-0.03	1.00	0.25	0.13	0.53
TIME_FOR_PASSION	0.04	-0.04	-0.10	0.25	1.00	0.16	0.51
WEEKLY_MEDITATION	0.17	-0.13	-0.08	0.13	0.16	1.00	0.38
WORK_LIFE_BALANCE_SCORE	0.20	-0.24	-0.28	0.53	0.51	0.38	1.00

Obviando la correlación entre una variable consigo misma, cuyo valor es, obviamente, 1, se puede observar que la correlación máxima positiva que se obtiene es la existente entre las variables WORK_LIFE_BALANCE_SCORE y ACHIEVEMENT, con un valor positivo de 0.58. La correlación inversa más grande obtenida se encuentra entre las variables WORK_LIFE_BALANCE_SCORE y DAILY_STRESS, con un valor de -0.37. Estos resultados iniciales aportan la siguiente información: a medida que una persona obtiene un mayor número de logros anuales memorables (tanto personales como profesionales) presumiblemente la puntuación final de la encuesta será mayor, sin tener en cuenta el resto de variables. Por el contrario, y de igual forma que en el caso anterior, si una persona experimenta unos niveles de estrés altos, se esperará una menor puntuación.

A partir de esta lectura inicial, el análisis descriptivo girará en torno a las variables que mayor coeficiente de correlación han obtenido, es decir, WORK_LIFE_BALANCE_SCORE y ACHIEVEMENT.

5. Cálculo del resumen del dataset original para observar valores como los valores máximos y mínimos de cada variable, los cuartiles, la media, etc.

```
> summary(db)
  Timestamp          FRUITS_VEGGIES   DAILY_STRESS   PLACES_VISITED   CORE_CIRCLE     SUPPORTING_OTHERS SOCIAL_NETWORK    ACHIEVEMENT       DONATION
 Length:3328        Min.   :0.000   Min.   :0.000   Min.   : 0.00   Min.   :  0.000   Min.   : 0.000    Min.   : 0.000   Min.   : 0.000   Min.   :0.000
 Class :character   1st Qu.:2.000   1st Qu.:2.000   1st Qu.: 2.00   1st Qu.:  3.000   1st Qu.: 3.000    1st Qu.: 4.000   1st Qu.: 2.000   1st Qu.:1.000
 Mode  :character   Median :3.000   Median :3.000   Median : 5.00   Median :  5.000   Median : 5.000    Median : 6.000   Median : 3.000   Median :3.000
                    Mean   :2.844   Mean   :2.834   Mean   : 5.14   Mean   :  5.308   Mean   : 5.404    Mean   : 6.483   Mean   : 3.855   Mean   :2.694
                    3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.: 8.00   3rd Qu.:  7.000   3rd Qu.: 9.000    3rd Qu.:10.000   3rd Qu.: 5.000   3rd Qu.:5.000
                    Max.   :5.000   Max.   :5.000   Max.   :10.00   Max.   :10.000    Max.   :10.000    Max.   :10.000   Max.   :10.000   Max.   :5.000
   BMI_RANGE      TODO_COMPLETED       FLOW        DAILY_STEPS      LIVE_VISION      SLEEP_HOURS     LOST_VACATION    DAILY_SHOUTING   SUFFICIENT_INCOME
 Min.   :1.000   Min.   : 0.00   Min.   : 0.000   Min.   : 1.000   Min.   :  0.000   Min.   : 1.000   Min.   : 0.00   Min.   : 0.000   Min.   :1.000
 1st Qu.:1.000   1st Qu.: 3.00   1st Qu.: 1.000   1st Qu.: 3.000   1st Qu.:  1.000   1st Qu.: 6.000   1st Qu.: 0.00   1st Qu.: 1.000   1st Qu.:1.000
 Median :1.000   Median : 6.00   Median : 2.000   Median : 5.000   Median :  3.000   Median : 7.000   Median : 0.00   Median : 2.000   Median :2.000
 Mean   :1.383   Mean   : 5.52   Mean   : 3.021   Mean   : 5.516   Mean   :  3.724   Mean   : 7.042   Mean   : 2.89   Mean   : 2.995   Mean   :1.705
 3rd Qu.:2.000   3rd Qu.: 8.00   3rd Qu.: 4.000   3rd Qu.: 8.000   3rd Qu.:  5.000   3rd Qu.: 8.000   3rd Qu.: 5.00   3rd Qu.: 5.000   3rd Qu.:2.000
 Max.   :2.000   Max.   :10.00   Max.   :10.000   Max.   :10.000   Max.   :  10.000   Max.   :10.000   Max.   :10.00   Max.   :10.000   Max.   :2.000
 PERSONAL_AWARDS  TIME_FOR_PASSION WEEKLY_MEDITATION      AGE              GENDER          WORK_LIFE_BALANCE_SCORE
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000    Length:3328        Length:3328        Min.   :480.0
 1st Qu.: 3.000   1st Qu.: 1.000   1st Qu.: 4.000    Class :character   Class :character   1st Qu.:633.9
 Median : 5.000   Median : 2.000   Median : 7.000    Mode  :character   Mode  :character   Median :662.6
 Mean   : 5.598   Mean   : 3.158   Mean   : 6.242                                          Mean   :663.6
 3rd Qu.: 8.000   3rd Qu.: 5.000   3rd Qu.:10.000                                          3rd Qu.:694.7
 Max.   :10.000   Max.   :10.000   Max.   :10.000                                          Max.   :818.3
```

6. Análisis descriptivo del subconjunto de datos (wlbs y achiev), que estudiará tanto las medidas de centralización como las de dispersión en ambas variables:

- **Medidas de centralización:**
    - Moda:

```
> moda_s
[1] 660.5
1129 Levels: 480 509.7 529.8 529.9 531 531.8 533.7 537.3 537.5 539.6 543.3 544.7 546.2 547.7 547.9 549 550.2 550.5 550.7 552 553 554.3 554.5 555.6 556.4 557.5 ... 818.3
> moda_a
[1] 2
Levels: 0 1 2 3 4 5 6 7 8 9 10
```

El valor más repetido en cuanto a puntuación total obtenida en la encuesta es 660.5, mientras que el valor más repetido para la variable Achievement es de 2 logros al año.

    - Media:

```
> media_s
[1] 663.5947
> media_a
[1] 3.855168
```

La puntuación media obtenida en la encuesta es de 663.59, mientras que el número de logros memorables anuales medio es de 3.86.

    - Mediana:

```
> mediana_s
[1] 662.6
> mediana_a
[1] 3
```

La mediana de los datos, en cuanto a la variable de puntuación total obtenida en la encuesta, se sitúa en el valor 662.6, mientras que la mediana de los logros anuales memorables está situada en el valor 3.

- **Medidas de dispersión:**
  - Cuartiles:

```
> cuartiles_s
   0%   25%   50%   75%  100%
480.0 633.9 662.6 694.7 818.3
> cuartiles_a
  0%  25%  50%  75% 100%
   0    2    3    5   10
```

    - Primer Cuartil (Q1):
      - Variable WORK_LIFE_BALANCE_SCORE: Q1 = 633.9
      - Variable ACHIEVEMENT: Q1 = 2 logros/año
    - Segundo Cuartil (Q2): equivalente a la mediana de los datos.
      - Variable WORK_LIFE_BALANCE_SCORE: Q2 = 662.6
      - Variable ACHIEVEMENT: Q2 = 3 logros/año
    - Tercer Cuartil (Q3):
      - Variable WORK_LIFE_BALANCE_SCORE: Q3 = 694.7
      - Variable ACHIEVEMENT: Q3 = 5 logros/año

  - Rango Intercuartílico:
    - IQR = Q3 – Q1

```
> Q1_s <- cuartiles_s[2]
> Q3_s <- cuartiles_s[4]
> IRQ_s <- Q3_s - Q1_s
> IRQ_s
 75%
60.8
> Q1_a <- cuartiles_a[2]
> Q3_a <- cuartiles_a[4]
> IRQ_a <- Q3_a - Q1_a
> IRQ_a
75%
  3
```

      - Variable WORK_LIFE_BALANCE_SCORE:
        - IRC = 60.8
      - Variable ACHIEVEMENT: IRC = 3 logros/año

  - Cuasivarianza:

```
> cuasivar_s
[1] 2037.538
> cuasivar_a
[1] 7.495098
```

La cuasivarianza, o varianza insesgada, aporta una estimación de la varianza de la población a partir de la muestra tomada (N-1 = 3328 – 1 = 3227), por lo que se puede deducir que la varianza para la variable WORK_LIFE_BALANCE_SCORE será próxima a 2037.538 de puntuación y, en el caso de la variable ACHIEVEMENT, el valor de ésta será próximo a 7.495 logros/año.

- Varianza:

```
> var_s
[1] 2036.926
> var_a
[1] 7.492846
```

La varianza poblacional de ambas variables han resultado ser próximas a los valores de las cuasivarianzas, como se esperaba y se ha calculado en el apartado anterior. Los resultados de las varianzas son 2036.926 de puntuación, para la variable WORK_LIFE_BALANCE_SCORE, y 7.493 logros/año para la variable ACHIEVEMENT.

- Cuasidesviación típica:

```
> cuasisd_s
[1] 45.1391
> cuasisd_a
[1] 2.737718
```

La cuasidesviación típica, de forma análoga a la cuasivarianza, aportará una estimación de lo que será la desviación típica de la población, pero en este caso calculada con la muestra tomada (3327 observaciones). En la variable WORK_LIFE_BALANCE_SCORE se ha obtenido un valor de 45.139 de puntuación, mientras que en el caso de la variable ACHIEVEMENT se ha obtenido un valor de 2.738 logros/año.

- Desviación típica:

```
> sd_s
[1] 45.13232
> sd_a
[1] 2.737306
```

Las desviaciones típicas obtenidas, como se esperaba por lo calculado anteriormente, ha resultado ser 45.132 de puntuación para la variable WORK_LIFE_BALANCE_SCORE, y de 2.737 logros/año para la variable ACHIEVEMENT. Estos valores reflejan cuán dispersos están los datos respecto de

la media de los mismos. Por tanto, se podría decir que los datos correspondientes a la variable WORK_LIFE_BALANCE_SCORE están más dispersos que los de la otra variable, dado que su desviación típica queda lejos de ser próxima a 0, mientras que la desviación típica de la variable ACHIVEMENT es más próxima a este valor.

- **Covarianza y correlación:**

Se define previamente un subconjunto de datos, denominado 'score_achiev', con los datos de las dos variables objeto de estudio: WORK_LIFE_BALANCE_SCORE y ACHIEVEMENT. Además, se almacenan los datos de cada variable en las respectivas variables de R, Y y X, respectivamente.

```
  #Covarianza y correlación:
score_achiev <- select(db2,WORK_LIFE_BALANCE_SCORE,ACHIEVEMENT)
X = as.numeric(db2$ACHIEVEMENT)
Y = as.numeric(db2$WORK_LIFE_BALANCE_SCORE)
```

- Covarianza muestral:

```
> cov_muestral <- cov * ((muestra-1)/muestra)
> cov_muestral
[1] 72.17171
```

La covarianza muestral es un valor estadístico que indica el grado de variación conjunta que existe entre dos variables aleatorias respecto a sus medias. En este caso concreto, el valor obtenido es positivo, lo que muestra que el comportamiento de los datos en ambas variables es similar, es decir, cuando una variable toma valores grandes, los valores de la otra variable se obtendrán también grandes y, de forma análoga, pasará también para valores pequeños. Además, este valor aporta una estimación de lo que será la covarianza poblacional.

- Covarianza poblacional:

```
> cov <- cov(X,Y)
> cov
[1] 72.1934
```

La covarianza poblacional es igual a 72.193, próxima al valor de la covarianza muestral. El signo obtenido en ambas covarianzas muestra la tendencia en la relación lineal que existe entre las variables WORK_LIFE_BALANCE_SCORE y ACHIEVEMENT.

- Correlación:

```
> correlacion
                          WORK_LIFE_BALANCE_SCORE ACHIEVEMENT
WORK_LIFE_BALANCE_SCORE                    1.00        0.58
ACHIEVEMENT                                0.58        1.00
```

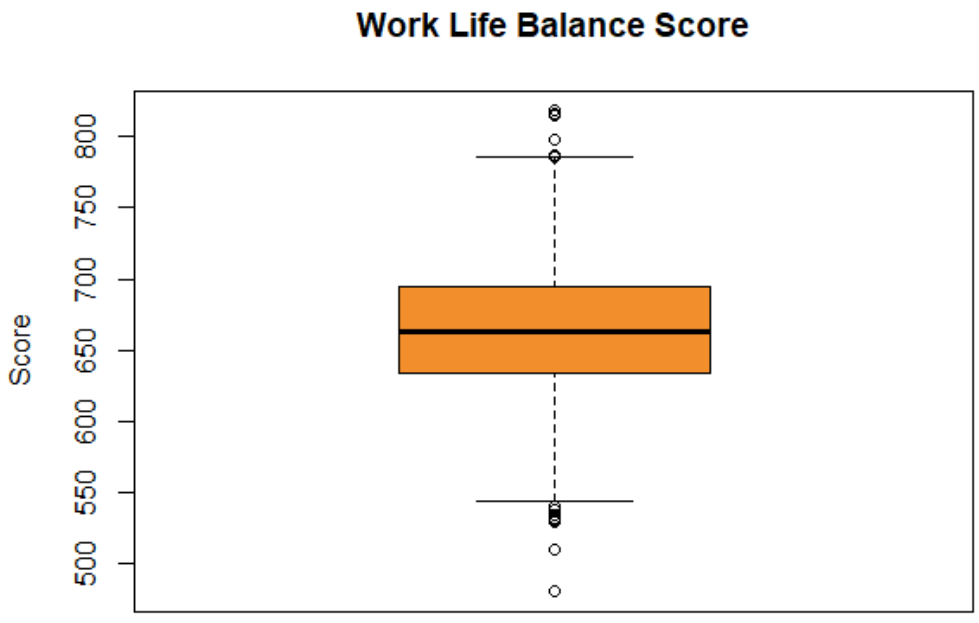
Como se puede comprobar, este valor de correlación positiva se corresponde con el valor obtenido inicialmente en la matriz de correlación de todas las variables. En este caso concreto, se obtiene un valor positivo de 0.58, por lo que se puede decir que existe una correlación moderada entre ambas variables, ya que este parámetro estadístico muestra la fuerza de la relación lineal entre ambas variables gracias a su magnitud. A medida que la persona tiene un número mayor de logros memorables anuales, se preveé una mayor puntuación final en la encuesta.

- **Gráfico de caja o boxplot:**

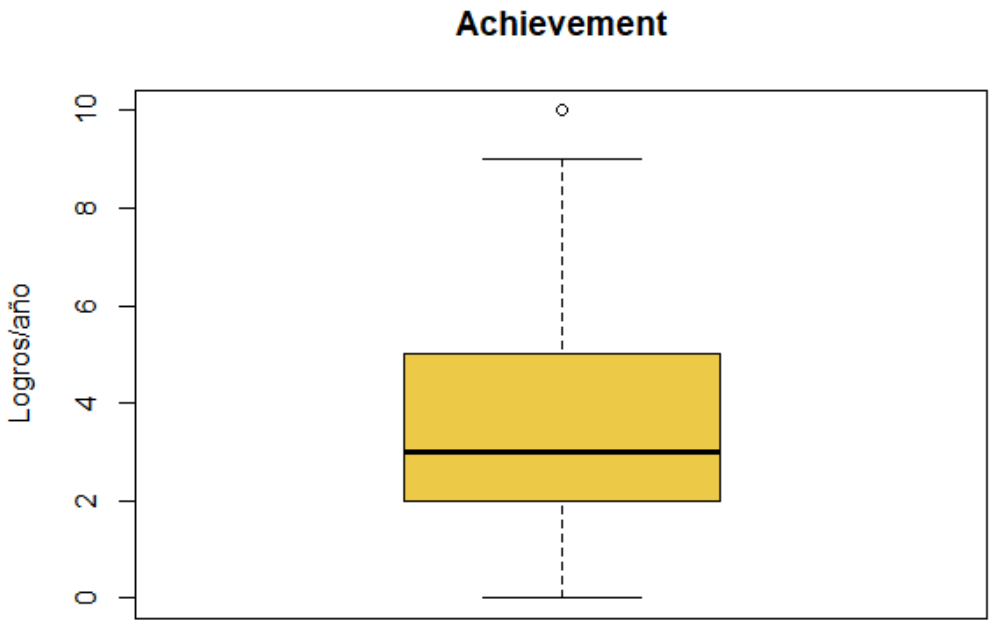
En este tipo de gráfico se representan los valores que permiten analizar descriptivamente los valores de la muestra de datos. En él se pueden ver los valores máximos y mínimos de la misma, la media, la mediana, los cuartiles, e incluso si existen outliers o valores fuera de rango. Además, según la dimensión de la caja, se puede saber si la muestra de datos es muy dispersa o no.

Individualmente, los gráficos obtenidos para sendas variables estudiadas, son los siguientes:

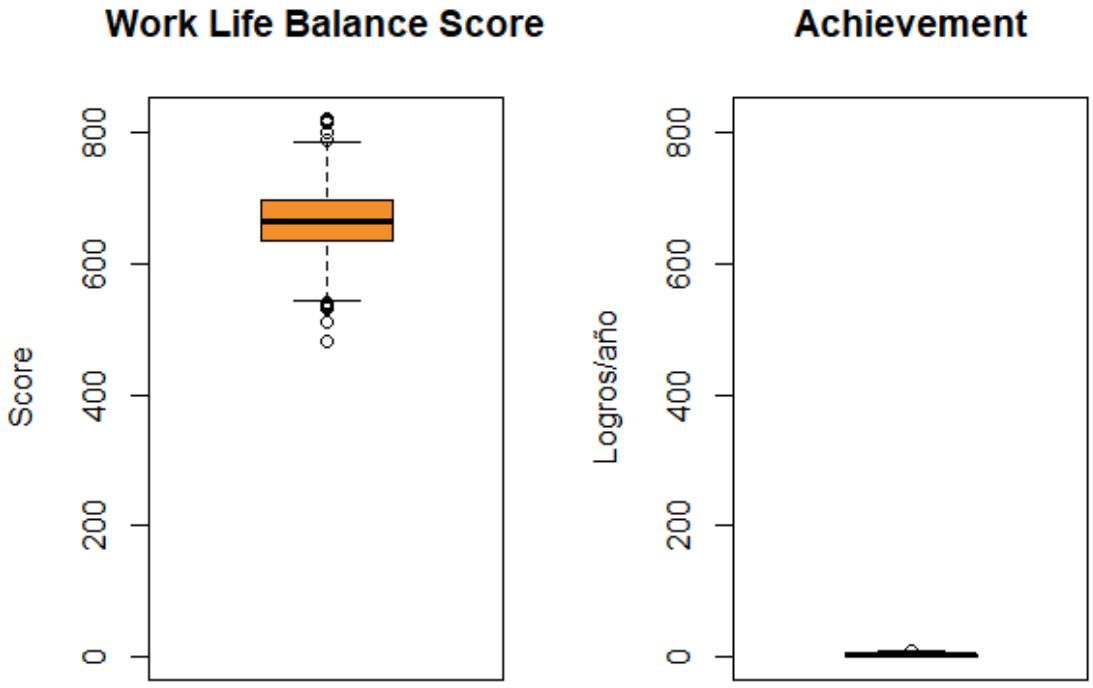
- Variable WORK_LIFE_BALANCE_SCORE:



- Variable ACHIEVEMENT:



Conjuntamente, se necesita escalar los ejes sobre los que se representan ambos gráficos, obteniendo el siguiente resultado:



Como se puede observar, tanto en los gráficos conjuntos como en los individuales, ambas variables cuentan con outliers. En el caso de la variable ACHIEVEMENT, se dispone de un solo valor atípico, mientras que en la otra variable se dispone de, al menos, 6 ó más de ellos. La muestra de datos de la variable WORK_LIFE_BALANCE_SCORE está mucho más dispersa que la de la otra variable, puesto que su caja tiene una mayor dimensión que la caja de ACHIEVEMENT, como también nos confirmaban los valores de sus desviaciones típicas. Además, esto también viene indicado por el rango o valores máximos y mínimos de cada conjunto de datos.

5. MODELO DE REGRESIÓN

Como el objetivo específico de esta práctica consiste en verificar que la puntuación total de la encuesta está basada en un modelo válido, se va a definir la variable WORK_LIFE_BALANCE_SCORE como la variable dependiente o explicada (variable Y), mientras que el resto de variables (excepto Timestamp) actuarán como variables independientes o explicativas. Esto significa que el modelo de regresión lineal a estimar será de tipo múltiple.

El procedimiento seguido para la estimación del modelo de regresión lineal es el siguiente:

1. Definición y asignación de datos a las diferentes variables que entrarán en el modelo. En su mayoría, se trata de variables numéricas, a excepción de las categóricas mencionadas en apartados anteriores de esta memoria: DAILY_STRESS, BMI_RANGE, TODO_COMPLETED, SUFFICIENT_INCOME, AGE y GENDER.

```
#ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE:
  #Definición y asignación de datos de las diferentes variables del modelo:
Y = as.numeric(db2$WORK_LIFE_BALANCE_SCORE)
X = as.numeric(db2$FRUITS_VEGGIES)
A = as.factor(db2$DAILY_STRESS)
B = as.numeric(db2$PLACES_VISITED)
C = as.numeric(db2$CORE_CIRCLE)
D = as.numeric(db2$SUPPORTING_OTHERS)
E = as.numeric(db2$SOCIAL_NETWORK)
F = as.numeric(db2$ACHIEVEMENT)
G = as.numeric(db2$DONATION)
H = as.factor(db2$BMI_RANGE)
I = as.factor(db2$TODO_COMPLETED)
J = as.numeric(db2$FLOW)
K = as.numeric(db2$DAILY_STEPS)
L = as.numeric(db2$LIVE_VISION)
M = as.numeric(db2$SLEEP_HOURS)
N = as.numeric(db2$LOST_VACATION)
O = as.numeric(db2$DAILY_SHOUTING)
P = as.factor(db2$SUFFICIENT_INCOME)
Q = as.numeric(db2$PERSONAL_AWARDS)
R = as.numeric(db2$TIME_FOR_PASSION)
S = as.numeric(db2$WEEKLY_MEDITATION)
T = as.factor(db2$AGE)
U = as.factor(db2$GENDER)
```

2. Cálculo y estimación del modelo de regresión lineal múltiple. Se va a estimar la variabilidad de la puntuación total obtenida en la encuesta respecto al resto de variables:

```
  #Modelo original, con todas las variables:
lm.fit = lm(Y~X+A+B+C+D+E+F+G+H+I+J+K+L+M+N+O+P+Q+R+S+T+U)
summary(lm.fit)
```

Los resultados obtenidos son los siguientes:

```
> lm.fit = lm(Y~X+A+B+C+D+E+F+G+H+I+J+K+L+M+N+O+P+Q+R+S+T+U)
> summary(lm.fit)

Call:
lm(formula = Y ~ X + A + B + C + D + E + F + G + H + I + J +
    K + L + M + N + O + P + Q + R + S + T + U)

Residuals:
       Min         1Q     Median         3Q        Max
-5.225e-12 -1.520e-13 -1.700e-14  1.030e-13  1.037e-10

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  5.442e+02  3.058e-13  1.779e+15   <2e-16 ***
X            3.400e+00  2.406e-14  1.413e+14   <2e-16 ***
A1          -3.400e+00  1.715e-13 -1.982e+13   <2e-16 ***
A2          -6.800e+00  1.685e-13 -4.036e+13   <2e-16 ***
A3          -1.020e+01  1.662e-13 -6.136e+13   <2e-16 ***
A4          -1.360e+01  1.740e-13 -7.816e+13   <2e-16 ***
A5          -1.700e+01  1.821e-13 -9.333e+13   <2e-16 ***
B            1.700e+00  1.085e-14  1.566e+14   <2e-16 ***
C            1.700e+00  1.266e-14  1.343e+14   <2e-16 ***
D            1.700e+00  1.204e-14  1.412e+14   <2e-16 ***
E            1.700e+00  1.183e-14  1.437e+14   <2e-16 ***
F            1.700e+00  1.442e-14  1.179e+14   <2e-16 ***
G            3.400e+00  1.929e-14  1.762e+14   <2e-16 ***
H2          -1.700e+01  6.884e-14 -2.469e+14   <2e-16 ***
I1           1.700e+00  2.087e-13  8.147e+12   <2e-16 ***
I2           3.400e+00  1.958e-13  1.737e+13   <2e-16 ***
I3           5.100e+00  1.848e-13  2.759e+13   <2e-16 ***
I4           6.800e+00  1.911e-13  3.558e+13   <2e-16 ***
I5           8.500e+00  1.792e-13  4.742e+13   <2e-16 ***
I6           1.020e+01  1.874e-13  5.444e+13   <2e-16 ***
I7           1.190e+01  1.792e-13  6.641e+13   <2e-16 ***
I8           1.360e+01  1.826e-13  7.446e+13   <2e-16 ***
I9           1.530e+01  2.149e-13  7.121e+13   <2e-16 ***
I10          1.700e+01  2.094e-13  8.120e+13   <2e-16 ***
J            1.700e+00  1.650e-14  1.031e+14   <2e-16 ***
K            1.900e+00  1.195e-14  1.589e+14   <2e-16 ***
L            1.700e+00  1.056e-14  1.610e+14   <2e-16 ***
M            1.900e+00  2.638e-14  7.202e+13   <2e-16 ***
N           -1.700e+00  8.888e-15 -1.913e+14   <2e-16 ***
O           -1.700e+00  1.255e-14 -1.355e+14   <2e-16 ***
P2           1.700e+01  7.358e-14  2.310e+14   <2e-16 ***
Q            1.700e+00  1.191e-14  1.428e+14   <2e-16 ***
R            1.700e+00  1.419e-14  1.198e+14   <2e-16 ***
S            1.700e+00  1.132e-14  1.502e+14   <2e-16 ***
T36 to 50   -9.369e-14  8.162e-14 -1.148e+00    0.251
T51 or more -8.457e-14  9.428e-14 -8.970e-01    0.370
TLess than 20 -8.163e-14  1.043e-13 -7.830e-01    0.434
UMale       -4.400e-14  6.896e-14 -6.380e-01    0.523
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.827e-12 on 3290 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 5.487e+28 on 37 and 3290 DF,  p-value: < 2.2e-16
```

Este modelo queda definido mediante la siguiente ecuación:

$$Y = Intercept + X*x_1 + A*x_2 + B*x_3 + C*x_4 + D*x_5 + E*x_6 + F*x_7 + G*x_8 + H*x_9 + I*x_{10} + J*x_{11} + K*x_{12} + L*x_{13} + M*x_{14} + N*x_{15} + O*x_{16} + P*x_{17} + Q*x_{18} + R*x_{19} + S*x_{20} + T*x_{21} + U*x_{22} =>$$

En el caso de que la persona que responde la encuesta tenga las siguientes características, la ecuación resultante del modelo estimado quedaría de la siguiente forma:

- Características:
    - Sexo masculino.
    - Edad comprendida entre 21 y 35 años.
    - Grado de satisfacción respecto a los ingresos percibidos para poder cubrir las necesidades básicas: insuficiente.
    - Grado de satisfacción en cuanto a cumplimiento de objetivos en una lista de tareas semanales: 7 sobre 10.
    - Índice de masa corporal de tipo 2: mayor a 25.
    - Nivel de estrés diario sufrido: 3 sobre 5 (estrés moderado).
- Ecuación resultante:

$$Y = 544.20 + 3.40*x_1 - 10.20*x_2 + 1.70*x_3 + 1.70*x_4 + 1.70*x_5 + 1.70*x_6 + 1.70*x_7 + 3.40*x_8 - 1.70*x_9 + 11.90*x_{10} + 1.70*x_{11} + 1.90*x_{12} + 1.70*x_{13} + 1.90*x_{14} - 1.70*x_{15} - 1.70*x_{16} + 0*x_{17} + 1.70*x_{18} + 1.70*x_{19} + 1.70*x_{20} + 0*x_{21} + -4.40*10^{-14}*x_{22}$$

Como se puede comprobar, algunas variables categóricas como AGE y SUFFICIENT_INCOME entran dentro del modelo con valores nulos. Esto es así porque las características iniciales del supuesto individuo que contestaría la encuesta, en ambas variables, no obtienen estimación según el modelo de regresión, por lo que entran dentro de la ecuación con valor 0. Así pues, ésta sería la ecuación mediante la cual se realizarían predicciones sujetas a las características establecidas. Si se quisiera modificar alguna de las mismas, bastaría con elegir el coeficiente estimado dentro de la variable que se está modificando, es decir, si en lugar de sexo masculino fuera sexo femenino, el valor de la variable U sería igual a 0. O, por ejemplo, si la persona indicara un grado de satisfacción de 5 sobre 10, en cuanto a cumplimiento de objetivos en una lista de tareas semanales, el valor de la variable I sería igual a 8.50.

Analizando los coeficientes de determinación, tanto ajustado como sin ajustar, se puede afirmar que se trata de un modelo de regresión perfecto, pues el valor es igual a 1. Esto quiere decir que el modelo queda explicado en un 100% por las variables que entran dentro de él.

Se podría decir que este resultado era de esperar inicialmente puesto que, a priori, se sabe que la puntuación que se obtiene en la encuesta viene parametrizada por los resultados obtenidos en las diferentes cuestiones abordadas.

A pesar de ello, analizando los resultados del p-valor, se puede observar que existen dos variables categóricas que no son significativas dentro del modelo: AGE y GENDER. Ambas variables obtienen un p-valor mayor al nivel de significanción, definido por defecto con un valor de 0.05 (5%), lo que quiere decir que la variabilidad de la puntuación obtenida en la encuesta no va a ser significativamente dependiente del género y la edad. Otra lectura de estos resultados sería la siguiente: al haber obtenido un p-valor mayor al nivel de significación, en ambas variables, no se puede rechazar la hipótesis nula que indica que las estimaciones de ambos parámetros son iguales a cero. Por ello, se propone la estimación del modelo eliminando estas variables, con el objetivo de comprobar si sufre grandes modificaciones y si, de igual forma, se pueden realizar predicciones con menos información.

En este caso concreto se está realizando la estimación del modelo con 3290 grados de libertad, por lo que si se dejaran fuera del mismo las dos variables no significativas, se estarían recuperando 4 grados de libertad.

3. Cálculo y estimación del modelo de regresión lineal múltiple sin las variables categóricas AGE y GENDER:

```
#Modelo sin las variables categóricas AGE y GENDER:
lm.fit2 = lm(Y~X+A+B+C+D+E+F+G+H+I+J+K+L+M+N+O+P+Q+R+S)
summary(lm.fit2)
```

Se define un nuevo modelo de regresión lineal múltiple, obteniendo los siguientes resultados:

```
> lm.fit2 = lm(Y~X+A+B+C+D+E+F+G+H+I+J+K+L+M+N+O+P+Q+R+S)
> summary(lm.fit2)

Call:
lm(formula = Y ~ X + A + B + C + D + E + F + G + H + I + J +
    K + L + M + N + O + P + Q + R + S)

Residuals:
       Min         1Q     Median         3Q        Max
-5.282e-12 -1.420e-13 -1.700e-14  9.900e-14  1.038e-10

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept)  5.442e+02  2.947e-13  1.846e+15   <2e-16 ***
X            3.400e+00  2.360e-14  1.441e+14   <2e-16 ***
A1          -3.400e+00  1.712e-13 -1.986e+13   <2e-16 ***
A2          -6.800e+00  1.675e-13 -4.060e+13   <2e-16 ***
A3          -1.020e+01  1.652e-13 -6.176e+13   <2e-16 ***
A4          -1.360e+01  1.728e-13 -7.870e+13   <2e-16 ***
A5          -1.700e+01  1.806e-13 -9.412e+13   <2e-16 ***
B            1.700e+00  1.078e-14  1.577e+14   <2e-16 ***
C            1.700e+00  1.257e-14  1.353e+14   <2e-16 ***
D            1.700e+00  1.184e-14  1.436e+14   <2e-16 ***
E            1.700e+00  1.164e-14  1.460e+14   <2e-16 ***
F            1.700e+00  1.436e-14  1.184e+14   <2e-16 ***
G            3.400e+00  1.899e-14  1.790e+14   <2e-16 ***
H2          -1.700e+01  6.748e-14 -2.519e+14   <2e-16 ***
I1           1.700e+00  2.073e-13  8.199e+12   <2e-16 ***
I2           3.400e+00  1.944e-13  1.749e+13   <2e-16 ***
I3           5.100e+00  1.837e-13  2.776e+13   <2e-16 ***
I4           6.800e+00  1.896e-13  3.587e+13   <2e-16 ***
I5           8.500e+00  1.777e-13  4.783e+13   <2e-16 ***
I6           1.020e+01  1.854e-13  5.501e+13   <2e-16 ***
I7           1.190e+01  1.770e-13  6.724e+13   <2e-16 ***
I8           1.360e+01  1.804e-13  7.537e+13   <2e-16 ***
I9           1.530e+01  2.125e-13  7.200e+13   <2e-16 ***
I10          1.700e+01  2.063e-13  8.240e+13   <2e-16 ***
J            1.700e+00  1.649e-14  1.031e+14   <2e-16 ***
K            1.900e+00  1.191e-14  1.596e+14   <2e-16 ***
L            1.700e+00  1.054e-14  1.613e+14   <2e-16 ***
M            1.900e+00  2.628e-14  7.231e+13   <2e-16 ***
N           -1.700e+00  8.877e-15 -1.915e+14   <2e-16 ***
O           -1.700e+00  1.245e-14 -1.365e+14   <2e-16 ***
P2           1.700e+01  7.275e-14  2.337e+14   <2e-16 ***
Q            1.700e+00  1.187e-14  1.432e+14   <2e-16 ***
R            1.700e+00  1.409e-14  1.206e+14   <2e-16 ***
S            1.700e+00  1.122e-14  1.515e+14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.827e-12 on 3294 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 6.156e+28 on 33 and 3294 DF,  p-value: < 2.2e-16
```

La ecuación resultante, en este caso, sería la siguiente:

$Y = Intercept + X*x_1 + A*x_2 + B*x_3 + C*x_4 + D*x_5 + E*x_6 + F*x_7 + G*x_8 + H*x_9 + I*x_{10} + J*x_{11} + K*x_{12} + L*x_{13} + M*x_{14} + N*x_{15} + O*x_{16} + P*x_{17} + Q*x_{18} + R*x_{19} + S*x_{20} =>$

Que, para exactamente las mismas características definidas en el caso anterior (sin conocer el sexo y edad de la persona que contesta la encuesta), quedaría de la siguiente forma:

$Y = 544.20 + 3.40*x_1 - 10.20*x_2 + 1.70*x_3 + 1.70*x_4 + 1.70*x_5 + 1.70*x_6 + 1.70*x_7 + 3.40*x_8 - 1.70*x_9 + 11.90*x_{10} + 1.70*x_{11} + 1.90*x_{12} + 1.70*x_{13} + 1.90*x_{14} - 1.70*x_{15} - 1.70*x_{16} + 0*x_{17} + 1.70*x_{18} + 1.70*x_{19} + 1.70*x_{20}$

Como se puede observar, se han obtenido exactamente las mismas estimaciones de todos los parámetros que entran dentro de la ecuación.

Si se analizan de nuevo los coeficientes de determinación, se puede observar que, a pesar de haber eliminado dos variables, se obtiene un modelo explicado al 100% por el resto de las variables. Esto quiere decir que no se ha obtenido penalización alguna en el modelo por la eliminación de las variables género y edad.

De forma análoga, se analizan los valores del p-valor respecto al nivel de significación, y se observa que todos ellos son menores a éste, lo que significa que la variabilidad de la puntuación obtenida en la encuesta es significativamente dependiente de la variabilidad del resto de variables.

En el apartado siguiente se desea analizar el mismo modelo de regresión pero eliminando por completo las variables categóricas, con el objetivo de analizar un modelo totalmente numérico y ver qué diferencias existen respecto a éste último.

4. Cálculo y estimación del modelo de regresión lineal múltiple sin las variables categóricas:

```
#Modelo sin ninguna variable categórica:
lm.fit3 = lm(Y~X+B+C+D+E+F+G+J+K+L+M+N+O+Q+R+S)
summary(lm.fit3)
```

Los resultados obtenidos son los siguientes:

```
> lm.fit3 = lm(Y~X+B+C+D+E+F+G+J+K+L+M+N+O+Q+R+S)
> summary(lm.fit3)

Call:
lm(formula = Y ~ X + B + C + D + E + F + G + J + K + L + M +
    N + O + Q + R + S)

Residuals:
    Min      1Q  Median      3Q     Max
-38.746  -8.817   0.553  10.470  32.552

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 532.04295    1.57554  337.69   <2e-16 ***
X             4.48510    0.16865   26.59   <2e-16 ***
B             2.34516    0.07668   30.58   <2e-16 ***
C             1.92186    0.09006   21.34   <2e-16 ***
D             1.51659    0.08518   17.80   <2e-16 ***
E             1.76770    0.08331   21.22   <2e-16 ***
F             2.17091    0.10275   21.13   <2e-16 ***
G             3.20007    0.13640   23.46   <2e-16 ***
J             2.09576    0.11801   17.76   <2e-16 ***
K             2.27004    0.08500   26.71   <2e-16 ***
L             2.01729    0.07532   26.78   <2e-16 ***
M             2.65135    0.18829   14.08   <2e-16 ***
N            -1.96180    0.06284  -31.22   <2e-16 ***
O            -2.59814    0.08472  -30.67   <2e-16 ***
Q             1.91120    0.08476   22.55   <2e-16 ***
R             1.77772    0.10129   17.55   <2e-16 ***
S             1.92203    0.08035   23.92   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.21 on 3311 degrees of freedom
Multiple R-squared:  0.9148,    Adjusted R-squared:  0.9144
F-statistic:  2221 on 16 and 3311 DF,  p-value: < 2.2e-16
```

La ecuación resultante, en este otro caso, sería la siguiente:

$Y = Intercept + X*x_1 + B*x_2 + C*x_3 + D*x_4 + E*x_5 + F*x_6 + G*x_7 + J*x_8 + K*x_9 + L*x_{10} + M*x_{11} + N*x_{12} + O*x_{13} + Q*x_{14} + R*x_{15} + S*x_{16} =>$

Las estimaciones de los parámetros han variado respecto al modelo anterior, por lo que para poder realizar predicciones (ahora sin tener en cuenta las características iniciales establecidas), se utilizaría la siguiente ecuación:

$Y = 532.043 + 4.49*x_1 + 2.35*x_2 + 1.92*x_3 + 1.52*x_4 + 1.77*x_5 + 2.17*x_6 + 3.20*x_7 + 2.10*x_8 + 2.27*x_9 + 2.02*x_{10} + 2.65*x_{11} - 1.96*x_{12} - 2.60*x_{13} + 1.91*x_{14} + 1.78*x_{15} + 1.92*x_{16}$

De igual forma que en el modelo anterior, el p-valor de todas las variables sigue siendo menor al nivel de significación, por lo que la variabilidad de la puntuación obtenida en la encuesta sigue siendo significativamente dependiente de la variabilidad del resto de

variables, con la salvedad de que, en este caso, el modelo queda peor explicado al disminuir los coeficientes de determinación del 100% al 91.5%. Esto quiere decir que se ha obtenido penalización por la eliminación de información.

5. Análisis de multicolinealidad, o dependencia lineal entre las diferentes variables que entran dentro del modelo de regresión, mediante el coeficiente GVIF (Generalized Variance Inlfation Factor). Éste permite analizar este fenómeno con variables categóricas y numéricas:

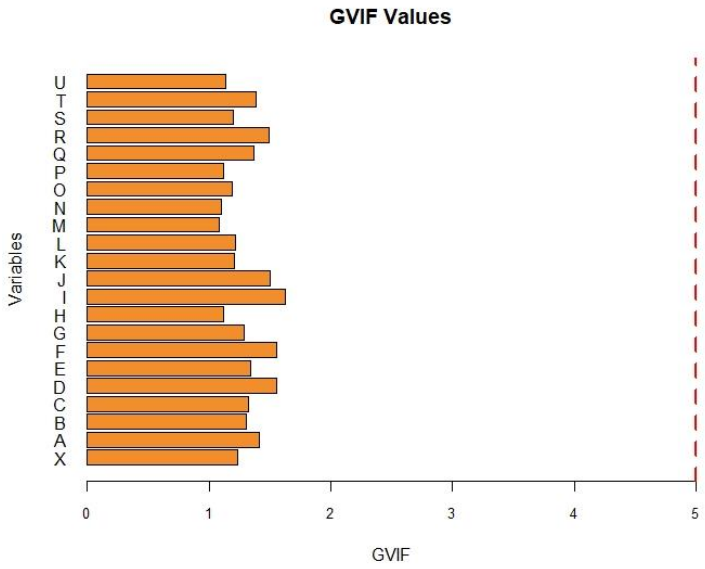
```
#ANÁLISIS DE MULTICOLINEALIDAD:
model <- lm(WORK_LIFE_BALANCE_SCORE~X+A+B+C+D+E+F+G+H+I+J+K+L+M+N+O+P+Q+R+S+T+U, data = db2)

gvif <- vif(model)
gvif_values <- gvif[,1]
barplot(gvif_values, main = "GVIF Values", xlab = "GVIF", ylab = "Variables",
        xlim = c(0,5), las = 1, cex.axis = 0.8, horiz = TRUE, col = "#F28E2B")

abline(v = 5, lwd = 3, lty = 2, col = 'red')
```

Los resultados obtenidos son los siguientes:

```
> gvif_values
       X        A        B        C        D        E        F        G        H        I        J
1.236970 1.416921 1.307753 1.322332 1.555112 1.339110 1.553650 1.288851 1.116386 1.630260 1.497312
       K        L        M        N        O        P        Q        R        S        T        U
1.209523 1.218830 1.087008 1.098444 1.189298 1.121770 1.367040 1.489997 1.197231 1.388581 1.135913
```

Conceptualmente puede parecer que algunas variables pueden estar fuertemente relacionadas con otras (DONATION frente a SUPPORTING_OTHERS), sin embargo, los valores que muestra este análisis son inferiores a 5, lo que indica que no existe multicolinealidad.

Gráficamente, se puede observar cuáles de las mismas pueden tener más o menos riesgo en este fenómeno, respecto al umbral que refleja la recta vertical en el valor 5:



6. Cálculo de los intervalos de confianza para cada variable estimada en el modelo de regresión. En este caso, se han obtenido todos los intervalos de confianza asociados a cada modelo de regresión calculado, obteniendo los siguientes resultados:

    - Intervalos de confianza para modelo de regresión lineal múltiple con todas las variables:

```
> confint(lm.fit)
                    2.5 %        97.5 %
(Intercept)   5.442000e+02  5.442000e+02
X             3.400000e+00  3.400000e+00
A1           -3.400000e+00 -3.400000e+00
A2           -6.800000e+00 -6.800000e+00
A3           -1.020000e+01 -1.020000e+01
A4           -1.360000e+01 -1.360000e+01
A5           -1.700000e+01 -1.700000e+01
B             1.700000e+00  1.700000e+00
C             1.700000e+00  1.700000e+00
D             1.700000e+00  1.700000e+00
E             1.700000e+00  1.700000e+00
F             1.700000e+00  1.700000e+00
G             3.400000e+00  3.400000e+00
H2           -1.700000e+01 -1.700000e+01
I1            1.700000e+00  1.700000e+00
I2            3.400000e+00  3.400000e+00
I3            5.100000e+00  5.100000e+00
I4            6.800000e+00  6.800000e+00
I5            8.500000e+00  8.500000e+00
I6            1.020000e+01  1.020000e+01
I7            1.190000e+01  1.190000e+01
I8            1.360000e+01  1.360000e+01
I9            1.530000e+01  1.530000e+01
I10           1.700000e+01  1.700000e+01
J             1.700000e+00  1.700000e+00
K             1.900000e+00  1.900000e+00
L             1.700000e+00  1.700000e+00
M             1.900000e+00  1.900000e+00
N            -1.700000e+00 -1.700000e+00
O            -1.700000e+00 -1.700000e+00
P2            1.700000e+01  1.700000e+01
Q             1.700000e+00  1.700000e+00
R             1.700000e+00  1.700000e+00
S             1.700000e+00  1.700000e+00
T36 to 50    -2.537191e-13  6.632823e-14
T51 or more  -2.694234e-13  1.002765e-13
TLess than 20 -2.861381e-13  1.228717e-13
UMale        -1.792221e-13  9.121159e-14
```

    - Intervalos de confianza para modelo de regresión lineal múltiple sin las variables categóricas AGE y GENDER:

```
> confint(lm.fit2)
              2.5 % 97.5 %
(Intercept) 544.2  544.2
X             3.4    3.4
A1           -3.4   -3.4
A2           -6.8   -6.8
A3          -10.2  -10.2
A4          -13.6  -13.6
A5          -17.0  -17.0
B             1.7    1.7
C             1.7    1.7
D             1.7    1.7
E             1.7    1.7
F             1.7    1.7
G             3.4    3.4
H2          -17.0  -17.0
I1            1.7    1.7
I2            3.4    3.4
I3            5.1    5.1
I4            6.8    6.8
I5            8.5    8.5
I6           10.2   10.2
I7           11.9   11.9
I8           13.6   13.6
I9           15.3   15.3
I10          17.0   17.0
J             1.7    1.7
K             1.9    1.9
L             1.7    1.7
M             1.9    1.9
N            -1.7   -1.7
O            -1.7   -1.7
P2           17.0   17.0
Q             1.7    1.7
R             1.7    1.7
S             1.7    1.7
```

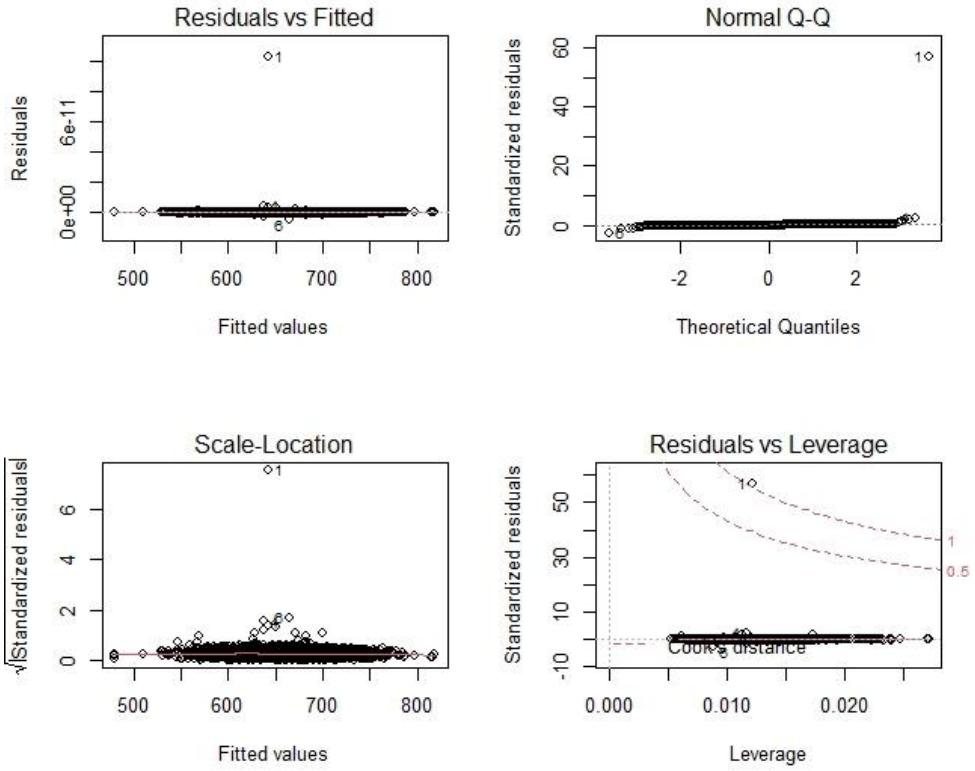
- Intervalos de confianza para modelo de regresión lineal múltiple sin las variables categóricas:

```
> confint(lm.fit3)
                  2.5 %     97.5 %
(Intercept) 528.953806 535.132085
X             4.154432   4.815761
B             2.194810   2.495509
C             1.745284   2.098442
D             1.349581   1.683601
E             1.604355   1.931045
F             1.969447   2.372376
G             2.932626   3.467505
J             1.864381   2.327142
K             2.103377   2.436697
L             1.869603   2.164972
M             2.282174   3.020519
N            -2.085017  -1.838588
O            -2.764260  -2.432026
Q             1.745014   2.077384
R             1.579124   1.976317
S             1.764493   2.079561
```

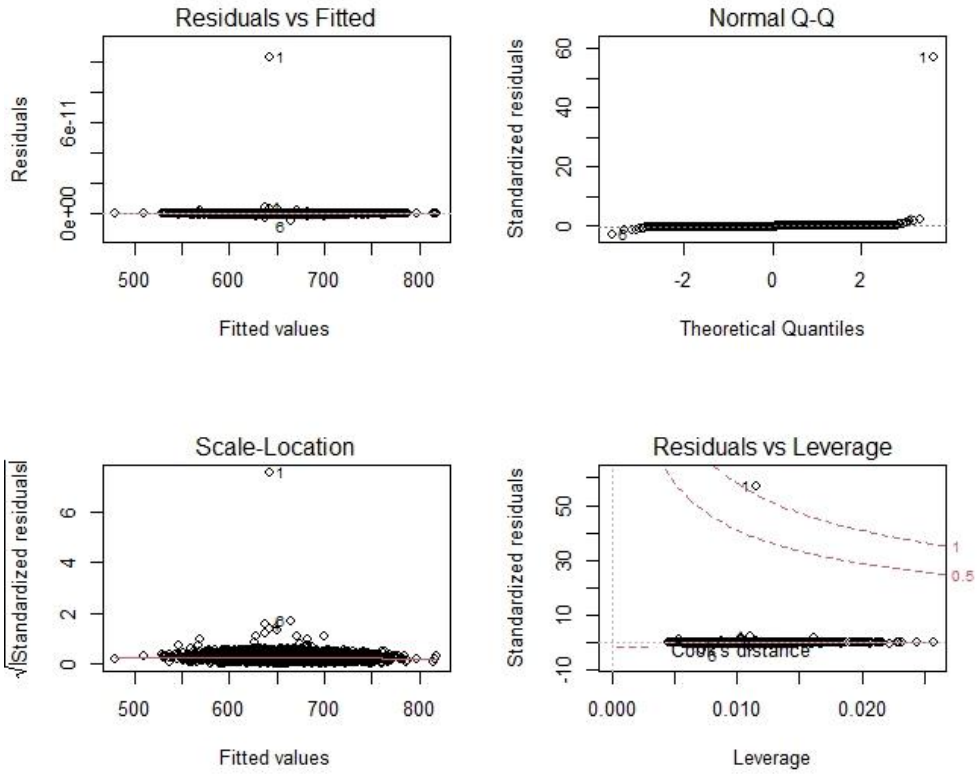
7. Obtención de gráficos de diagnóstico. De forma análoga al apartado anterior, se obtienen los gráficos de diagnóstico, con sus 4 subtipos de gráficos, para todos los modelos de regresión calculados:

```
#GRÁFICOS DE DIAGNÓSTICO:
par(mfrow=c(2,2))
plot(lm.fit)
plot(lm.fit2)
plot(lm.fit3)
```

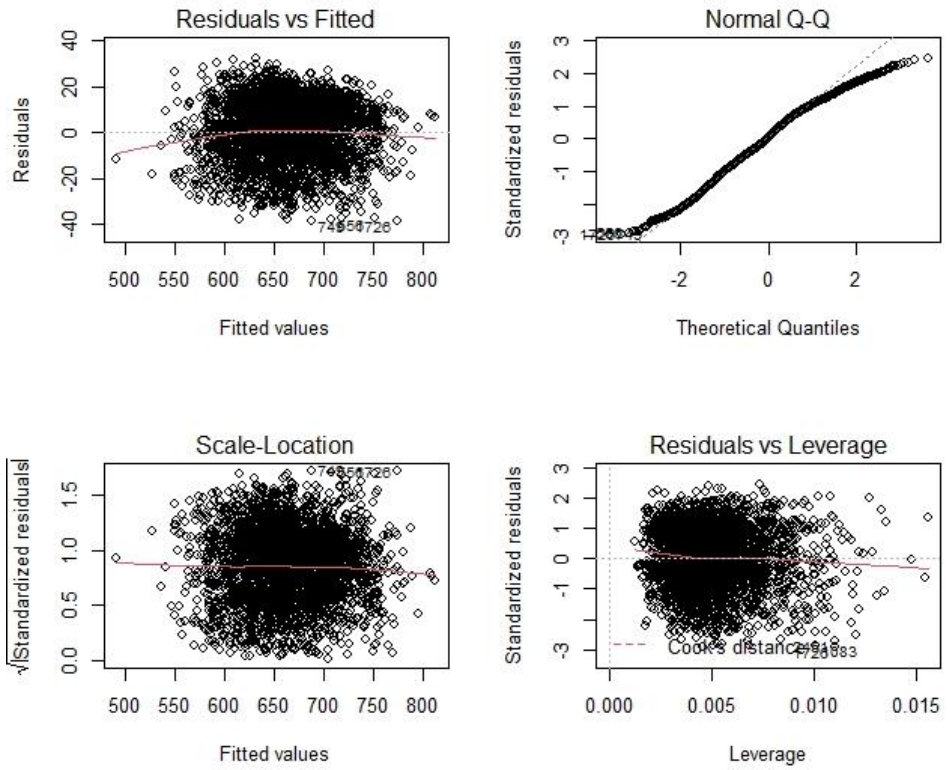
- Gráficos de diagnóstico del modelo de regresión lineal múltiple con todas las variables:



- Gráficos de diagnóstico para modelo de regresión lineal múltiple sin las variables categóricas AGE y GENDER:



- Gráficos de diagnóstico para modelo de regresión lineal múltiple sin las variables categóricas:



Los gráficos de diagnóstico permiten determinar si el modelo explica bien el patrón de los datos. De izquierda a derecha y de arriba abajo, lo que explican estas 4 gráficas es el comportamiento de los residuos en función de: si tienen patrones lineales (Residuals vs. Fitted), si se distribuyen normalmente (Normal Q-Q), si se distribuyen por igual a lo largo de los predictores (Scale-Location) o si existen observaciones influyentes (Residuals vs. Leverage).

En estos casos se puede observar que, para los dos primeros modelos, los resultados obtenidos son muy similares, mientras que para el tercer modelo existen diferencias notables. Los dos primeros casos muestran, en todos los tipos de gráficos, casos ideales, puesto que están basados en modelos muy bien explicados. Los datos se ajustan a lo que parece ser una recta, mientras que en el tercer modelo esa recta sufre una pequeña deformación, curvándose levemente.

8. Predicciones.

   En este apartado se han utilizado las respuestas ofrecidas por dos personas voluntarias: un varón y una mujer, ambos con edad comprendida entre los 21 y 35 años. Se pretende predecir la puntuación obtenida de cada uno de los individuos para cada modelo estimado. Para ello, se exponen a continuación los valores de las respuestas ofrecidas por sendos individuos y, a continuación, las estimaciones obtenidas:

VARIABLE	TRADUCCIÓN	RESPUESTA		UNIDADES
		Varón	Mujer	
FRUITS_VEGGIES	Frutas_Verduras	2	1	Piezas/día (0-5)
DAILY_STRESS	Estres_Diario	3	4	Puntuación (0-5)
PLACES_VISITED	Lugares_Visitados	6	6	Lugares/año (0-10)
CORE_CIRCLE	Circulo_Personal	10	8	Personas (0-10)
SUPPORTING_OTHERS	Ayudar_Otros	10	9	Personas (0-10)
SOCIAL_NETWORK	Redes_Sociales	10	10	Personas/día (0-10)
ACHIEVEMENT	Retos_Memorables	5	6	Logros/año (0-10)
DONATION	Donacion_Causas	2	1	Veces/año (0-5)
BMI_RANGE	Indice_Masa_Corporal	2	1	Puntuación (1-2)
TODO_COMPLETED	Lista_Tareas	7	8	Puntuación (0-10)
FLOW	Estado_Inmersion	6	5	Horas/día (0-10)
DAILY_STEPS	Pasos_Diarios	10	3	Miles de pasos/día (0-10)
LIVE_VISION	Vision_Vida	10	10	Años (0-10)
SLEEP_HOURS	Horas_Sueño	7	7	Horas/noche (1-10)
LOST_VACATION	Vacaciones_Perdidas	1	0	Días/año (0-10)
DAILY_SHOUTING	Dias_Gritando	0	0	Veces/semana (0-10)
SUFFICIENT_INCOME	Ingresos_Basicos	1	2	Puntuación (1-2)
PERSONAL_AWARDS	Reconocimientos_Personales	7	8	Reconocimientos/vida (0-10)
TIME_FOR_PASSION	Tiempo_Pasiones	3	1	Horas/día (0-10)
WEEKLY_MEDITATION	Meditacion_Semanal	3	2	Veces/semana (0-10)
AGE	Edad	21 to 35		Rangos
GENDER	Genero	M	F	M-F

Definición de parámetros que se sustituirán en las ecuaciones de cada modelo:

```
#CÁLCULO DE PREDICCIONES:
  #Datos varón 21-35 años:
newdata_lm = data.frame(X=2,A=as.factor(3),B=6,C=10,D=10,E=10,F=5,G=2,H=as.factor(2),I=as.factor(7),
                        J=6,K=10,L=10,M=7,N=1,O=0,P=as.factor(1),Q=7,R=3,S=3,T=as.factor("21 to 35"),
                        U=as.factor("Male"))

newdata_lm2 = data.frame(X=2,A=as.factor(3),B=6,C=10,D=10,E=10,F=5,G=2,H=as.factor(2),I=as.factor(7),
                         J=6,K=10,L=10,M=7,N=1,O=0,P=as.factor(1),Q=7,R=3,S=3)

newdata_lm3 = data.frame(X=2,B=6,C=10,D=10,E=10,F=5,G=2,J=6,K=10,L=10,M=7,N=1,O=0,Q=7,R=3,S=3)

  #Datos mujer 21-35 años:
data_lm = data.frame(X=1,A=as.factor(4),B=6,C=8,D=9,E=10,F=6,G=1,H=as.factor(1),I=as.factor(8),
                     J=5,K=3,L=10,M=7,N=0,O=0,P=as.factor(2),Q=8,R=1,S=2,T=as.factor("21 to 35"),
                     U=as.factor("Female"))

data_lm2 = data.frame(X=1,A=as.factor(4),B=6,C=8,D=9,E=10,F=6,G=1,H=as.factor(1),I=as.factor(8),
                      J=5,K=3,L=10,M=7,N=0,O=0,P=as.factor(2),Q=8,R=1,S=2)

data_lm3 = data.frame(X=1,B=6,C=8,D=9,E=10,F=6,G=1,J=5,K=3,L=10,M=7,N=0,O=0,Q=8,R=1,S=2)
```

Comandos utilizados en RStudio para predicción en los 3 modelos estimados:

```
  #Predicciones:
    #Varón:
predict.lm(lm.fit,newdata_lm,interval="confidence")
predict.lm(lm.fit2,newdata_lm2,interval="confidence")
predict.lm(lm.fit3,newdata_lm3,interval="confidence")

    #Mujer:
predict.lm(lm.fit,data_lm,interval="confidence")
predict.lm(lm.fit2,data_lm2,interval="confidence")
predict.lm(lm.fit3,data_lm3,interval="confidence")
```

Los resultados obtenidos para el varón de 21 a 35 años de edad son los siguientes:

```
> predict.lm(lm.fit,newdata_lm,interval="confidence")
    fit   lwr   upr
1 692.1 692.1 692.1
> predict.lm(lm.fit2,newdata_lm2,interval="confidence")
    fit   lwr   upr
1 692.1 692.1 692.1
> predict.lm(lm.fit3,newdata_lm3,interval="confidence")
       fit      lwr      upr
1 720.9234 719.1304 722.7164
```

El primer y segundo modelo demuestran, como cabía esperar, que se obtiene la misma puntuación en los mismos intervalos de confianza: 692.1 puntos sobre 1000. Esto es así porque, como se ha mencionado con anterioridad, ambos modelos son explicados en el mismo porcentaje (100%) a pesar de estar estimados con un número de variables diferente.

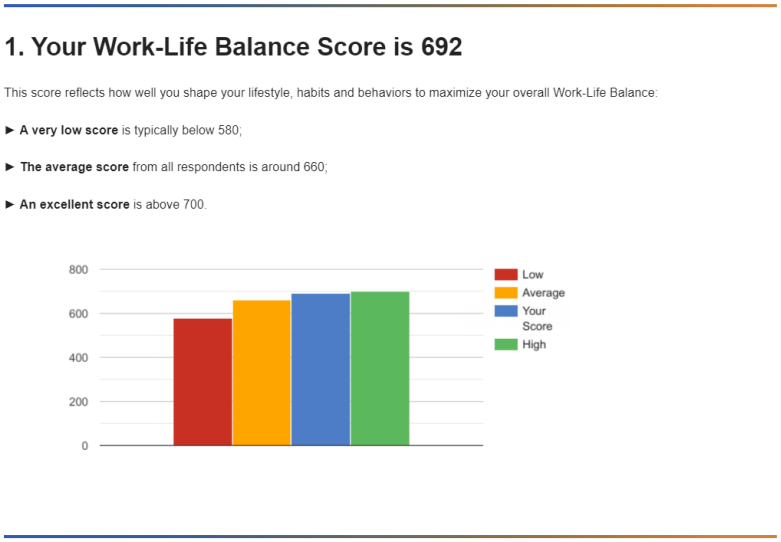
El tercer modelo, sin embargo, presenta diferencias respecto a los otros dos, con unos intervalos de confianza mayores: 720.92 puntos sobre 1000, aunque el resultado real podría oscilar entre los 719.13 y los 722.72 puntos.

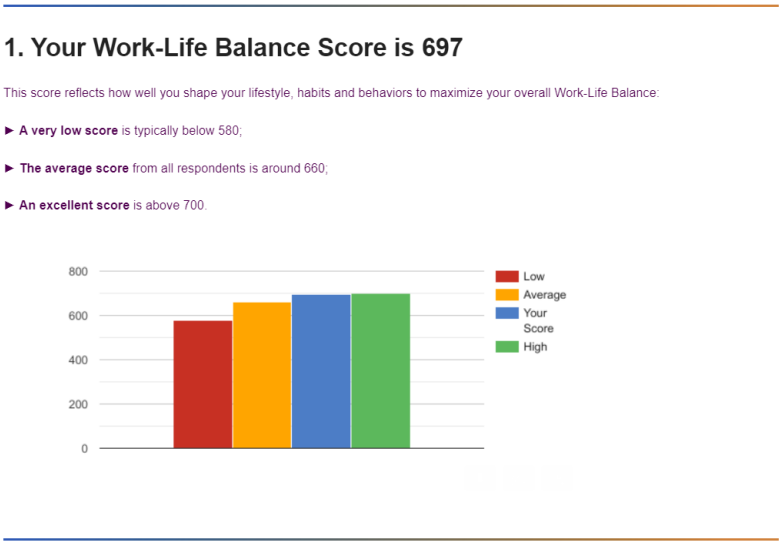
Para la mujer con mismo rango de edad los resultados son los siguientes:

```
> predict.lm(lm.fit,data_lm,interval="confidence")
    fit   lwr   upr
1 697.5 697.5 697.5
> predict.lm(lm.fit2,data_lm2,interval="confidence")
    fit   lwr   upr
1 697.5 697.5 697.5
> predict.lm(lm.fit3,data_lm3,interval="confidence")
       fit      lwr      upr
1 690.4583 688.5489 692.3678
```

Como se puede corroborar, análogamente al apartado anterior, los dos primeros modelos presentan el mismo resultado: 697.5 puntos sobre 1000. Mientras que si se estimara mediante el tercer modelo, se obtendría una puntuación diferente: 690.46 sobre 1000, aunque el resultado real podría oscilar entre los 688.55 y los 692.37 puntos.

Además, como comprobación adicional, se han realizado las encuestas de ambos individuos, con el objetivo de comprobar que se van a obtener cerca de los 692 puntos para el varón y los 697 para la mujer. Los resultados de la misma, respectivamente, son los siguientes:





En el apartado Anexos, además, se adjuntan los dos informes recibidos por el/la organizador/a de la encuesta, donde se explica con mayor detalle la puntuación y lo que refleja en los hábitos de las personas encuestadas.

6. CONCLUSIONES

La realización previa de una matriz de correlación permite observar las diferentes relaciones existentes entre los datos obtenidos, lo que facilita poder establecer un objetivo fundamentado en una relación más o menos fuerte respecto al resto de datos.

Los resultados obtenidos en el análisis del modelo de regresión en el que se basa la variable WORK_LIFE_BALANCE_SCORE pueden resultar obvios porque, aunque no se conoce el procedimiento de cálculo de la misma, se sabe que la naturaleza de los datos que recoge esta variable dependen directamente de las preguntas que contesta el usuario.

La ventaja de analizar un resultado que, a priori, parece obvio, es la verificación del mismo y posible confirmación del modelo de regresión. Además, saber que el modelo en el que están basados los datos es un modelo perfecto, permite poder configurarlo de tal forma que se puedan obtener estimaciones con el mismo nivel de confianza pero a partir de otros parámetros o información, es decir, estimaciones de las puntuaciones con la misma fiabilidad pero con menor información.

7. TRABAJOS FUTUROS, LIMITACIONES Y MEJORAS

Respecto a las **limitaciones y mejoras del dataset** se pueden destacar 2 de ellas.

Para poder trabajar con el dataset se debe hacer una suposición inicial que resulta ser la primera limitación detectada, y es la siguiente: no se conocen las dimensiones reales de la base de datos. Esto quiere decir, si se han cogido los datos correspondientes al año 2016, se sabe que este subconjunto consta de 3328 observaciones, pero no se sabe si esas observaciones se han realizado a personas diferentes o no. Una misma persona ha podido realizar la encuesta en diferentes momentos dentro del mismo año. Por ello, se ve necesario establecer como premisa inicial el hecho de que cada observación se considere como única, como si cada una de ellas correspondiera a un individuo diferente.

La mejora propuesta sería implementar un mecanismo, en la propia encuesta, que permita detectar si se trata de la misma persona o no. Podría ser, por ejemplo, a través del correo electrónico al que se envían los resultados. Este apartado no es obligatorio a la hora de realizar la encuesta, pero si se estableciera la obligatoriedad, se podría detectar qué usuario ha utilizado un mismo correo en más de una ocasión. Esto no implica que se pueda eliminar la limitación completamente, puesto que una persona puede tener N correos diferentes, pero al menos permite estudiar también la evolución de los datos de esa persona a lo largo del año o años en los que ha realizado la encuesta.

La segunda limitación detectada es que el subconjunto de datos escogido del dataset inicial contiene información poco relacionada entre sí, como se puede comprobar en la matriz de correlación, por lo que si se quisiera fijar un objetivo específico diferente al establecido en esta práctica, sería necesaria la recaudación de más información de otro tipo de variables que permitan explicar mejor la variable a estudiar. Por ejemplo, si se quisiera estudiar alguna de las variables que reflejan los comportamientos de las personas basados en la dimensión cuidado del cuerpo, como por ejemplo conocer la variabilidad de ingesta de frutas y verduras de una persona (a través de la variable FRUITS_VEGGIES), sería necesario obtener información como:

- Cuántas y qué tipo de frutas y verduras hay en cada temporada del año.
- Disponibilidad de las frutas y verduras según origen de procedencia (oferta, demanda, precios, etc.).
- Problemática en cuanto a cosechas (por fenómenos meteorológicos, ambientales, biológicos, etc.)

- Número de enfermedades diagnosticadas al individuo (por la posible recomendación médica de mejora en la alimentación).
- Ocupación de la persona (puede que un deportista de élite o un entrenador personal tenga un estilo de vida más saludable, con la consecuente mayor ingesta de este tipo de alimentos que cualquier otra persona)

En la recolecta e inclusión de este tipo de información, se deberá tener especial cuidado en las variables que puedan ser dependientes unas de otras (multicolinealidad).

Como trabajos futuros dentro de este análisis se propone el estudio de los datos respecto a al tiempo, con el objetivo de observar las posibles tendecias que pueden tener los mismos, ya que el dataset original dispone de datos comprendidos en un período de tiempo de al menos 5 años (desde 2015 hasta 2021).

8. ANEXOS

Para la elaboración de esta práctica se han utilizado los siguientes documentos anexos a la memoria:

- Ficheros de texto en los que se recogen, por una parte el dataset original, y por otra parte el dataset sobre el que se ha realizado el correspondiente análisis estadístico:
    - Dataset original: Wellbeing_and_lifestyle_data_Kaggle.csv
    - Dataset utilizado: Dataset_Datos_2016.txt:
- Fichero R, donde se han elaborado todos los análisis estadísticos contemplados en esta memoria.
- Ficheros PDF con las puntuaciones finales proporcionadas por la página web de la encuesta:
    - Life_Satisfaction_Report for undefined_Female.pdf
    - Life_Satisfaction_Report for undefined_Male.pdf

9. REFERENCIAS

La bibliografía empleada como apoyo para la realización de esta práctica es la siguiente:

- "Manual" de R para consulta de funciones, tipos de datos, gráficos, etc.:

  Bosco Mendoza Vega, Juan. *R para principiantes*. Online.

  https://bookdown.org/jboscomendoza/r-principiantes4/

- Enlace web para consultar las diferentes paletas de color que existen en Rstudio:

  Javier. 2020. *Paletas de Colores en R*. Online.

  https://estadisticamente.com/paletas-de-colores-en-r/

- Enlace web para consulta de matrices de correlación:

  Salazar, Camila. 2018. *Cálculo de correlaciones*. Online.

  https://rpubs.com/camilamila/correlaciones

- Enlace web para consulta de configuración y/o personalización de gráficos en R:

  R Coder. 2021. *R Charts*. Online.

  https://r-charts.com/es/r-base/ejes/

- Enlace web para consulta de modelos de regresión y representación:

  Delgado, Ronald. 2018. *Introducción a los Modelos de Regresión en R*. Online.

  https://rpubs.com/rdelgado/395717

- Enlace web para consulta de modelos de regresión lineal múltiple:

  Data política. 2020. *Cómo predecir con Regresión Lineal Múltiple en RStudio (lm)*. Online.

  https://www.youtube.com/watch?v=XtDQD25Ejkc