TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS -- PRAC 2

ALUMNO: JOSÉ LUIS MARTÍN RAMOS

Práctica 2 (35% nota final)

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas. Para hacer esta práctica tendréis que trabajar en grupos de 2 personas.

Tendréis que entregar un solo archivo con el enlace Github (https://github.com) donde se encuentren las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos que corresponden a vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github. Aunque no se trata del mismo enunciado, los siguientes ejemplos de ediciones anteriores os pueden servir como guía:

- Ejemplo: https://github.com/Bengis/nba-gap-cleaning
- Ejemplo complejo (archivo adjunto).

Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en Tipología y ciclo de vida de los datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Descripción de la Práctica a realizar

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien cualquier dataset libre disponible en Kaggle (https://www.kaggle.com).

Algunos ejemplos de dataset con los que podéis trabajar son:

- Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)
- Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)

El último ejemplo corresponde a una competición activa de Kaggle de manera que, opcionalmente, podéis aprovechar el trabajo realizado durante la práctica para entrar en esta competición.

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y **justificar**) son las siguientes:

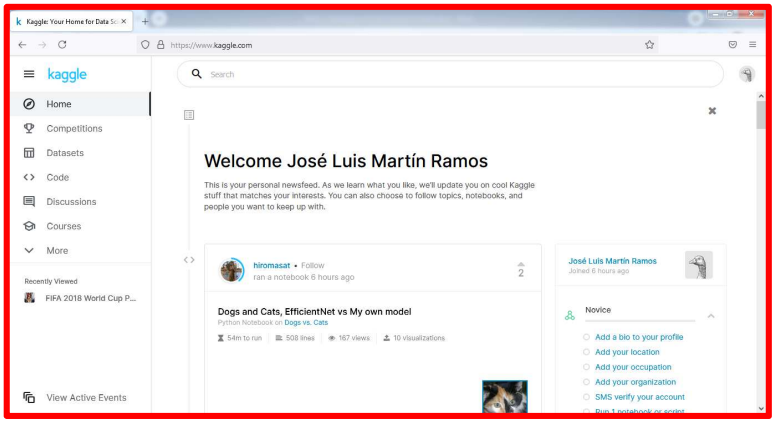
Ejercicio 1

Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

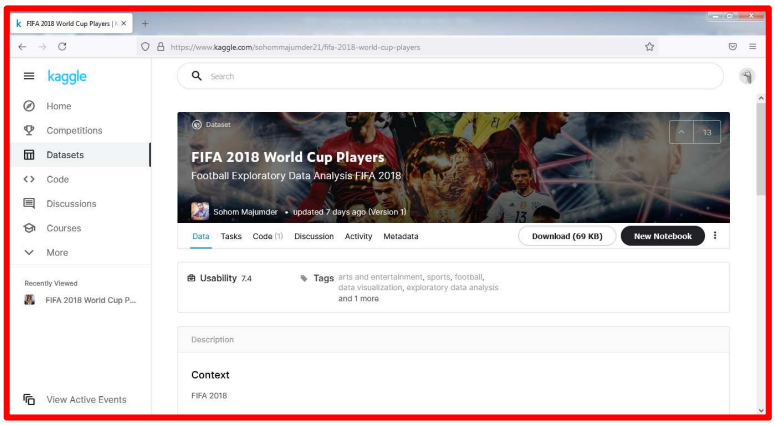
Respuesta

He elegido este dataset, por varias razones. La primera es que el compañero con el que realicé la primera práctica no debía sufrir las consecuencias de que mi situación laboral me haya impedido acometer esta práctica con más tiempo y por lo tanto la tengo que realizar solo (supongo y espero que él haya resuelto el problema de buscar compañero con tiempo). En segundo lugar, en año de Eurocopa, hacer un trabajo sobre fútbol me parecía interesante.

Para ellos he optado por seleccionar un dataset de los propuestos en el enunciado.

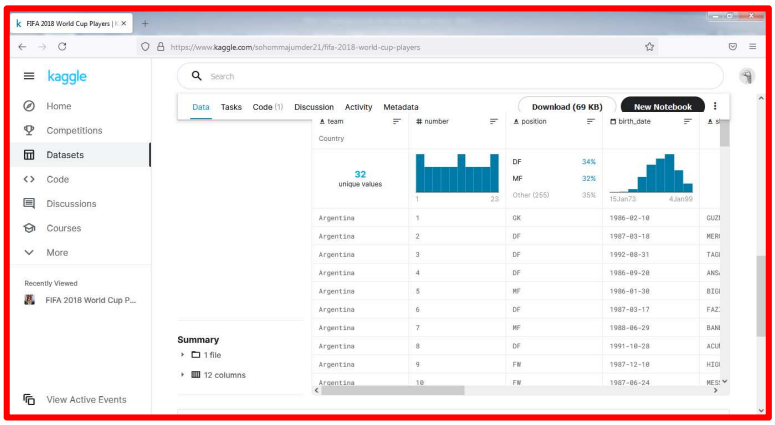


He elegido el de la “Copa del Mundo de 2018”.



Como entrenador de fútbol, nunca he tenido la oportunidad de responder a preguntas que en ocasiones salen en nuestro entorno sobre la importancia que tiene en un torneo de este tipo, cuestiones como: la envergadura de los jugadores (altura "height" y peso "weight"), la experiencia que tienen como internacionales "caps" o la liga en la que están participando. En muchas ocasiones se dice que las ligas española, inglesa, italiana o alemana están por encima del resto. Vamos a ver si la posición de los equipos en la clasificación (resultados obtenidos) puede tener algún tipo de relación con el número de jugadores de cada equipo en estas ligas.

El dataset disponible es el siguiente:



Ejercicio 2

Integración y selección de los datos de interés a analizar.

Respuesta

Los datos a analizar proceden del dataset descargado en la fase anterior. Es un dataset muy sencillo, no se observan a simple vista errores y no parece muy complejo de entender. Los datos que se manejan son los siguientes:

Team: Equipo (Argentina, Brasil, Colombia,…, Tunez y Uruguay).

Number: Número que lleva el jugador durante el torneo (de 1 a 23).

Position: Posición en el campo (GK "portero", DF "defensa", MF "centrocampista", FW "delantero") de su traducción de inglés a español (GK "Goal keaper", etc.)

Birth_date: Fecha de nacimiento (en formato año-mes-día AAAA-MM-DD).

Shirt_name: Nombre en la camiseta (Rojo, Otamendi,…)

Club: Equipo en el que juega antes del mundial (Manchester United FC,…)

Height: Altura en centímetros.

Weight: Peso en kilogramos.

Age: Edad.

Name: Nombre.

Caps: Internacionalidades.

Para leer este fichero con R, procedemos a descargarlo de la página y lo leemos con las siguientes líneas de código.

```
library(readr)
```

```
DatosIniciales <- read_csv
("C:/ESTRATEGICO/UO Catalunya/2021 Tipología y ciclo de vida de los datos/PRAC 2/all_wc_18_players_fifa.csv")
View(DatosIniciales)
```

Ejercicio 3.1

Limpieza de los datos. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Respuesta

Un breve resumen de los datos, nos lleva a ver que en el fichero tenemos 736 registros en 12 variables. Los datos que obtenemos son los siguientes:

```
     team              number          position          birth_date          shirt_name
 Length: 736        Min. : 1        Length: 736        Min. : 1973-01-15   Length: 736
 Class: character   1st Qu.: 6      Class: character   1st Qu.: 1987-09-20 Class: character
 Mode: character    Median: 12      Mode: character    Median: 1990-09-24  Mode: character
                    Mean: 12                           Mean: 1990-07-22
                    3rd Qu.: 18                        3rd Qu.: 1993-03-25
                    Max.: 23                           Max.: 1999-01-04
     club           height          weight             league              age
 Length: 736        Min.: 165.0     Min.: 59.00        Length: 736         Min.: 19.44
 Class: character   1st Qu.: 178.0  1st Qu.: 72.00     Class: character    1st Qu.: 25.22
 Mode: character    Median: 183.0   Median: 77.00      Mode: character     Median: 27.72
                    Mean: 182.4     Mean: 77.19                            Mean: 27.89
                    3rd Qu.: 187.0  3rd Qu.: 82.00                        3rd Qu.: 30.73
                    Max.: 201.0     Max.: 99.00                            Max.: 45.41
     name           caps
 Length: 736        Min.: 0.00
 Class: character   1st Qu.: 10.00
 Mode: character    Median: 25.00
                    Mean: 35.33
                    3rd Qu.: 52.00
                    Max.: 158.00
```

Tal como se comentó al principio, no se observan datos ausentes ni atípicos. La fecha de nacimiento concuerda con la edad, siendo el más joven un jugador de 19,44 años y el más viejo uno de 45,41 años. Respecto a la altura, tenemos jugadores de 165 a 201 centímetros y en el peso, jugadores de 59 a 99 kilogramos.

Un aspecto importante es el número de internacionalidades, hay jugadores que debutan en una competición de este tipo con “0” y otros que ya llevan “158” partidos con su selección.

En este caso no tenemos valores ausentes ni datos que puedan considerarse errores de grabación o mecanización. Si se hubieran observado, y una vez comprobado que no podemos obtener en valor correcto, tendríamos que optar por alguna de las siguientes opciones:

- Realizar una imputación automática. Podemos optar pos distintas opciones.

  - Imputar la media del valor (si procede) o cualquier otro estadístico de resumen adecuado (mediana, moda, …)
  - Imputar un valor obtenido por un algoritmo (regresión, etc.) teniendo en cuenta otras variables relacionadas.
- Descartar esos registros reduciendo el tamaño del dataset y perdiendo información válida para el registro en otras variables.
- Recodificar el error a N/A para que los algoritmos ignoren esos registros.

Ejercicio 3.2

Limpieza de los datos. Identificación y tratamiento de valores extremos.

Respuesta

Para determinar si una variable tiene valores atípicos, tenemos que observar los valores de la variable en el primer cuartil (Q1) y tercer cuartil (Q3). Entre Q1 y Q3 sabemos que están el 50% de los valores obtenidos en el estudio. A esta distancia se le llama **rango intercuantílico** (IQR: InterQuantile Range).

Se define como **valor atípico leve** aquel que dista 1,5 veces el rango intercuantílico por debajo de Q1 o por encima de Q3

$$q < Q1 - 1{,}5 \cdot IQR \text{ o bien } q > Q3 + 1{,}5 \cdot IQR$$

y **valor atípico extremo** aquel que dista 3 veces el rango intercantílico por debajo de Q1 o por encima de Q3

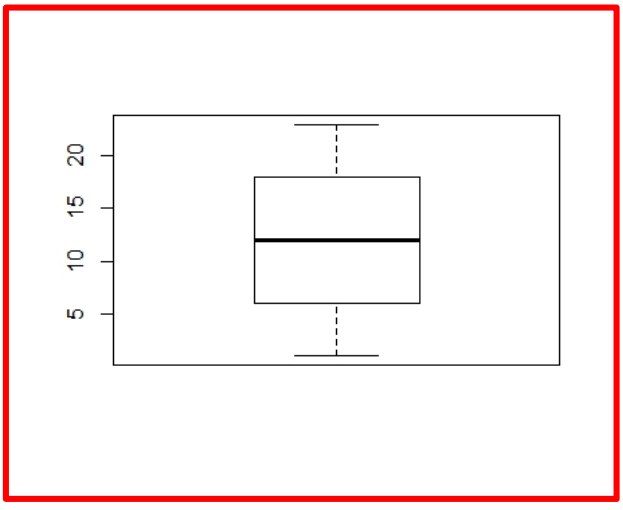
$$q < Q1 - 3 \cdot IQR \text{ o bien } q > Q2 + 3 \cdot IQR$$

Si sospecháramos que podemos tener valores atípicos por arriba o por debajo, podemos calcular cual serían los umbrales.

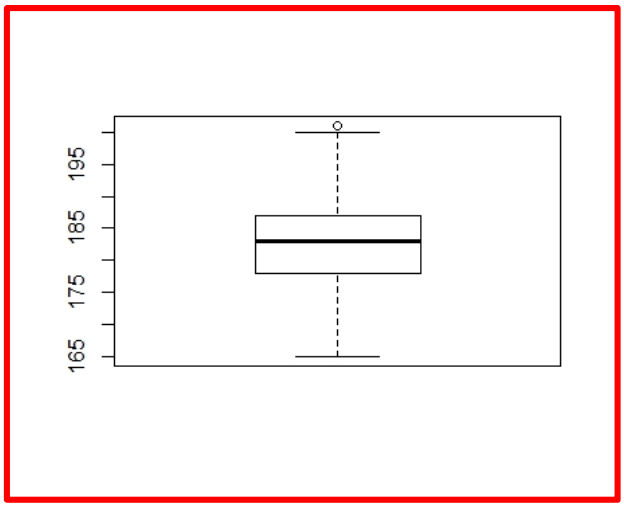
Podemos programas estos cálculos de forma sencilla teniendo en cuenta los datos obtenidos de os estadísticos descriptivos básicos.

Otra forma de obtener estos datos, es mediante el diagrama de tallos y hojas. Los vemos para las distintas variables numéricas del ejemplo.

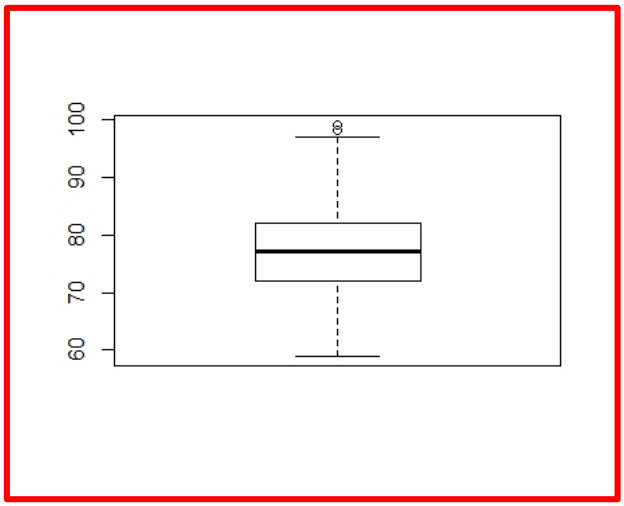
Para la variable **number**, lo calculamos aunque en este caso no tiene sentido: Observamos que no hay valores atípicos. Podríamos mejorar el gráfico poniendo etiquetas a los ejes y otras opciones disponibles en R.



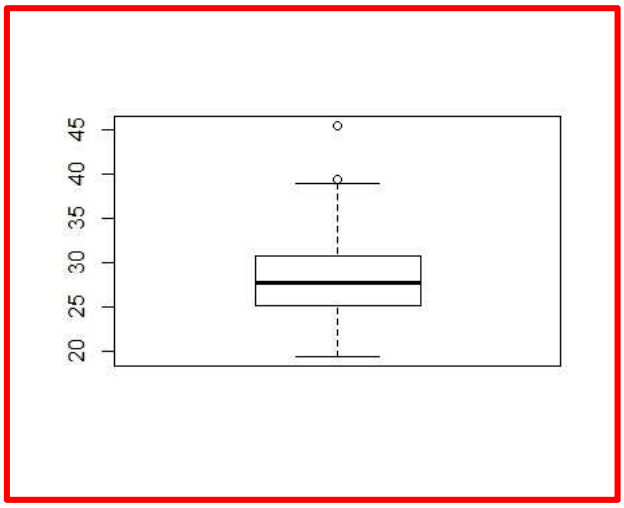
Para la variable **height**: Podemos comprobar que existe un valor atípico que se representa con el círculo superior. En este caso, se trata de una altura de 201 centímetros. Aunque aparezca como atípico, lo vamos a conservar por ser un valor correcto.



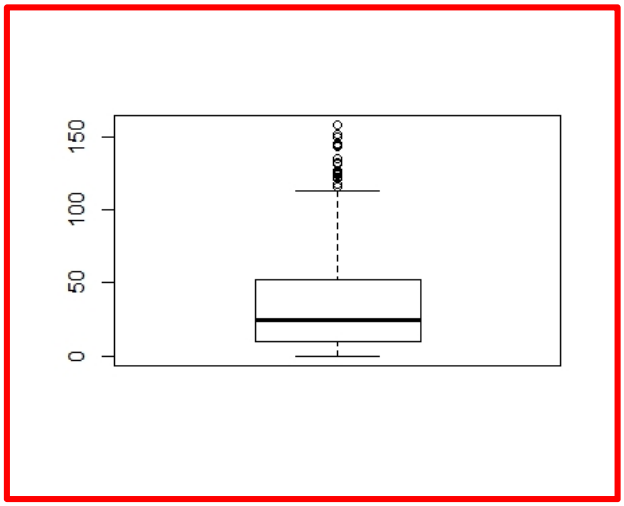
Para la variable **weight**: Observamos dos valores atípicos. En este caso el máximo es 99 kilogramos y hay varios de 98. Aunque pueda parecer raro, los observamos y son valores correctos que coinciden con jugadores de 188, 200, 193 y 178 centímetros. Los consideramos correctos.



Para la variable **age**: Nos aparece como atípico un valor de 45,41 y otro de 39,33. En ambos casos son valores correctos. El primero se trata de un jugador egipcio que destaca sobre los demás. El segundo está próximo a otros jugadores con estas edades. No es habitual, pero si ocurre en estos campeonatos que equipos no punteros recurren a jugadores con experiencia (por edad). Los consideramos correctos.



Para la variable **caps**: Observamos un número importante de jugadores con un numero de internacionalidades superiores a 100. Los vamos a considerar correcto porque se corresponden con jugadores de selecciones que juegan muchos partidos y son jugadores importantes para sus selecciones. Se trata de jugadores como Essan El-Hadary (egipcio de 45,41 años), Sergio Ramos en España o Cristiano Ronaldo en Portugal.



Ejercicio 4.1

Análisis de los datos. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Respuesta

Realizamos un análisis estadístico de las variables numéricas. Para ello obtenemos los estadísticos básicos y las frecuencias. También podemos obtener un gráfico de barras.

```
str(Datos$height) num [1:736] 192 181 169 181 175 199 175 172 184 170 ...
```

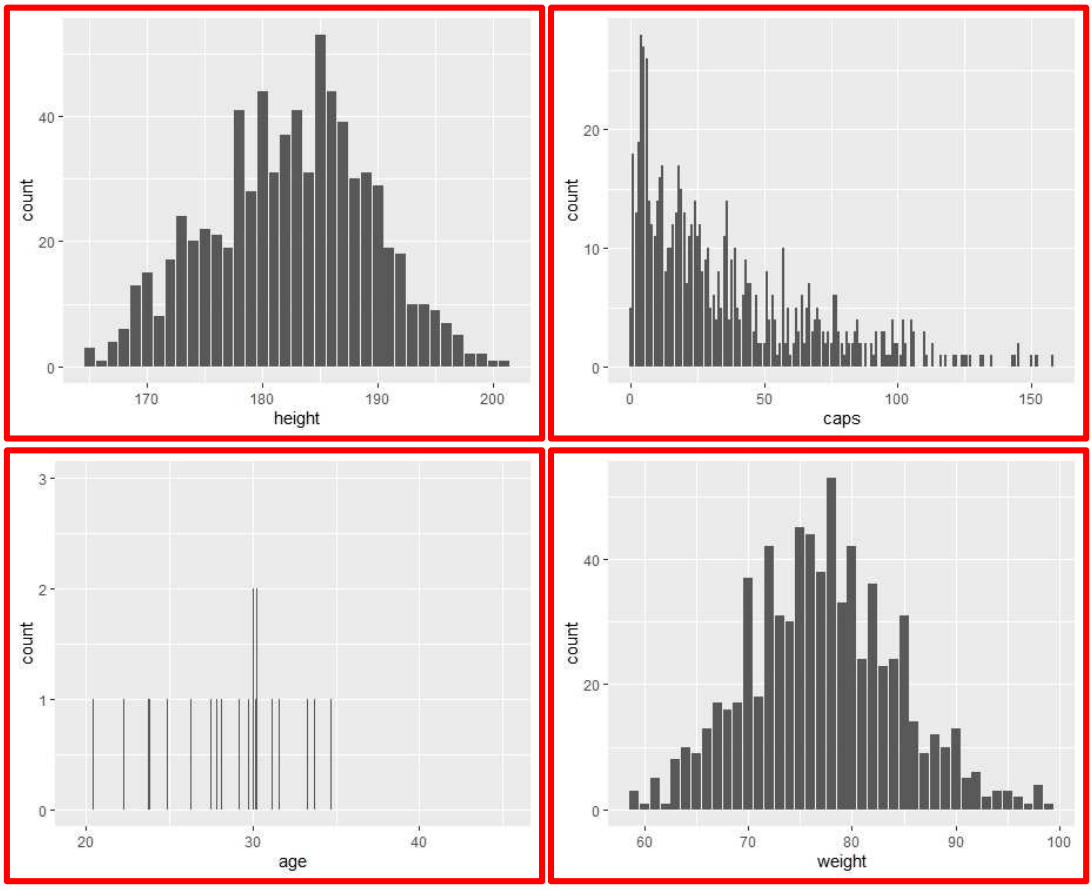
> summary(Datos$height)	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	165.0	178.0	183.0	182.4	187.0	201.0

> table(Datos$height)

165	166	167	168	169	170	…	198	199	200	201
3	1	4	6	13	15	…	2	2	1	1

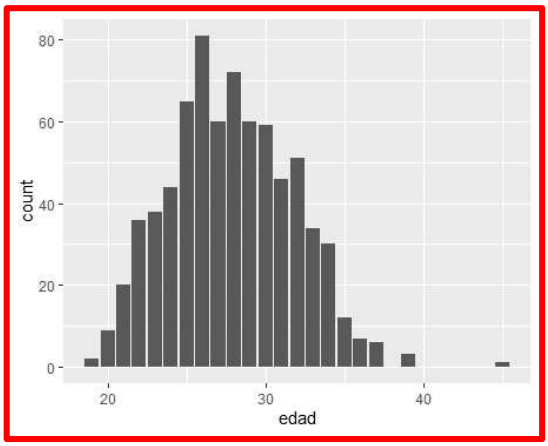
El resto de variables proporcionan sus propios datos.

Vamos a ver el histograma de estas variables.



Como podemos comprobar, las variables peso y altura siguen una distribución aparentemente normal, la variable internacionalidades no vamos a poder mantener esa normalidad y la variable edad, observamos que al estar calculada como numérica de tipo real, presenta muchos decimales. Tenemos que redondear al año completo u obtener desde la fecha de nacimiento una edad en años y meses (por ejemplo).

La nueva variable ya tiene una distribución más o menos normal donde podemos observar el dato atípico de 45 años.



Con estos datos vamos a calcular la correlación entre variables, a obtener una función de regresión y para algunos equipos, vamos a ver si hay diferencias entre las edades de algunas selecciones (igual lo hacemos para jugadores por continentes).

Ejercicio 4.2

Análisis de los datos. Comprobación de la normalidad y homogeneidad de la varianza.

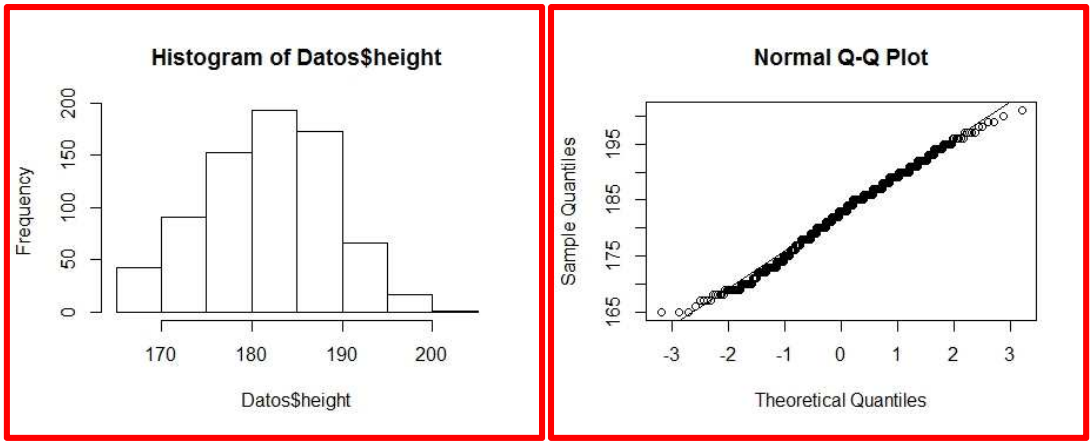
Respuesta

Hemos observado que los datos se ajustan a una distribución normal. Pero esto lo hemos hecho desde una perspectiva simple de observación de los histogramas.

Aplicamos ahora los test de normalidad y homogeneidad de la varianza a las cuatro variables de interés.

Para la variable **altura (height)**: Obtenemos un par de gráficos que nos sirven para ver cómo se presentan los datos. A continuación, con la función correspondiente, obtenemos los estadísticos de contraste de normalidad.

En el gráfico QQ de la derecha, observamos como las observaciones se ajustan correctamente al a línea diagonal de la distribución de los datos normales.



> skewness(Datos$height) [1] **-0.1215751**

> agostino.test(Datos$height) D'Agostino skewness test

data: Datos$height

skew = **-0.12158**, z = -1.35507, p-value = **0.1754**

alternative hypothesis: data have a skewness

Esta es una alternativa para calcular la normalidad (es el uso del test de asimetría). En este caso con un valor del estadítico de -0.1215751 y un p-valor de 0.1754, nada se opone a aceptar la normalidad de la variable. En este caso en realidad estamos comprobando si la variable es simétrica, pero es una alternativa.

Vemos ahora otra opción para la variable **peso (weight)**: Podemos utilizar el test de Shapiro-Wilk.

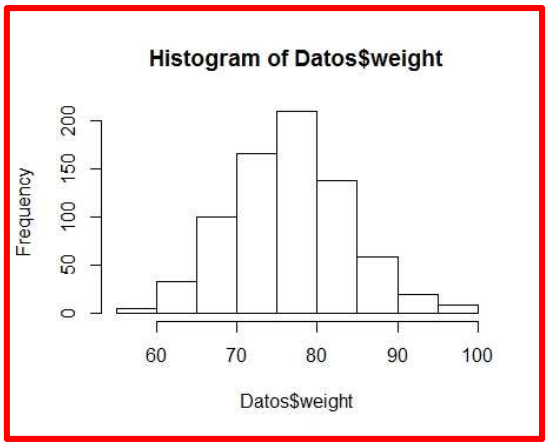
> print(x.test) Shapiro-Wilk normality test

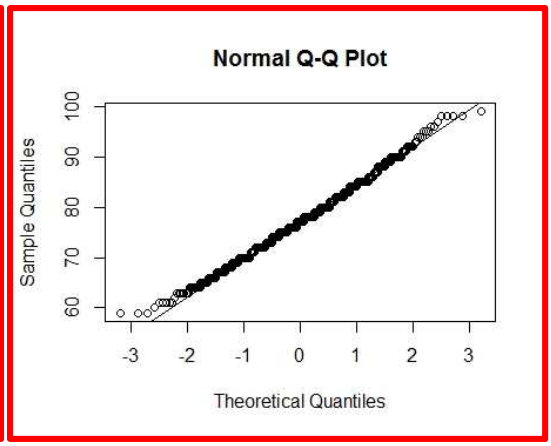
data: Datos$weight

W = 0.99432, p-value = **0.007408**

Con un p-valor de 0,0074 tendríamos que rechazar la hipótesis de normalidad para la variable peso.

Debajo podemos ver el histograma y el gráfico QQ.





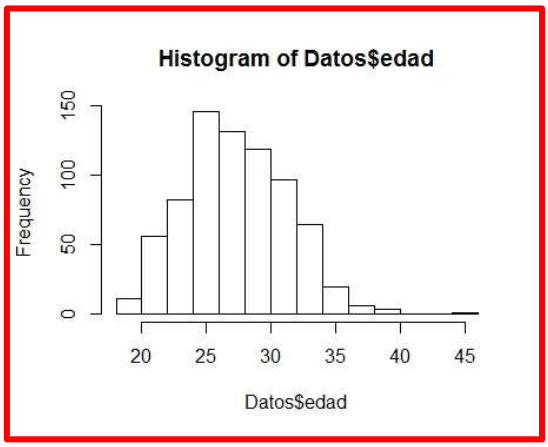
Para la **edad**: Realizando el test de S-W, obtenemos el siguiente resultado.

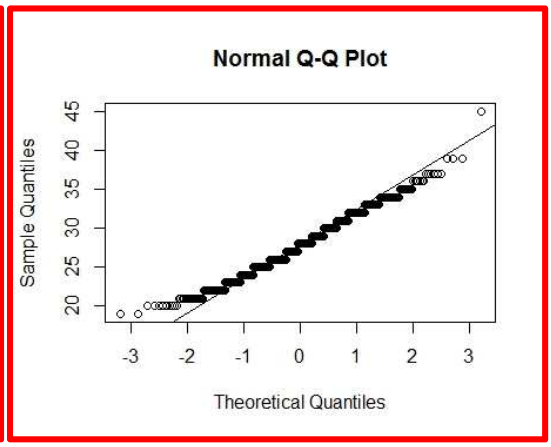
```
> print(x.test) Shapiro-Wilk normality test

data:  Datos$edad

W = 0.9852, p-value = 8.775e-07
```

Con un p-valor de 0,000 rechazamos la hipótesis de normalidad. También podemos verlo en el siguiente gráfico.





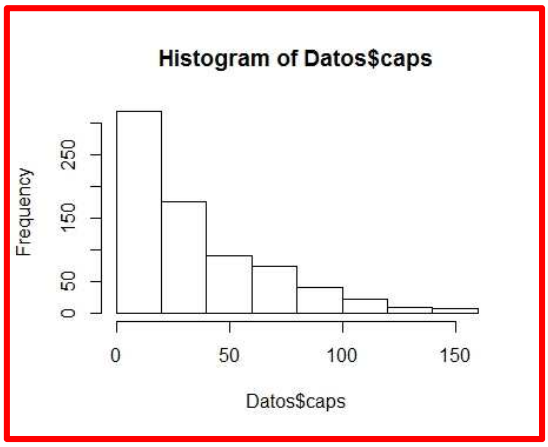
Y por último para la variable **internacionalidades**:

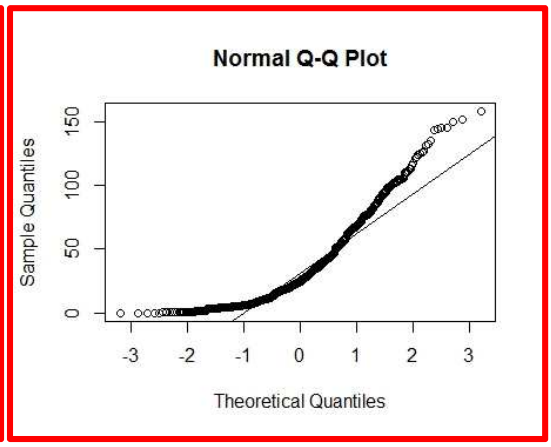
```
> print(x.test) Shapiro-Wilk normality test

data:  Datos$caps

W = 0.87487, p-value < 2.2e-16
```

El test de S-W con un p-valor de 0,000 nos obliga a rechazar la hipótesis de normalidad. En los gráficos siguientes se ve esta circunstancia claramente.





Ejercicio 4.3

Análisis de los datos. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

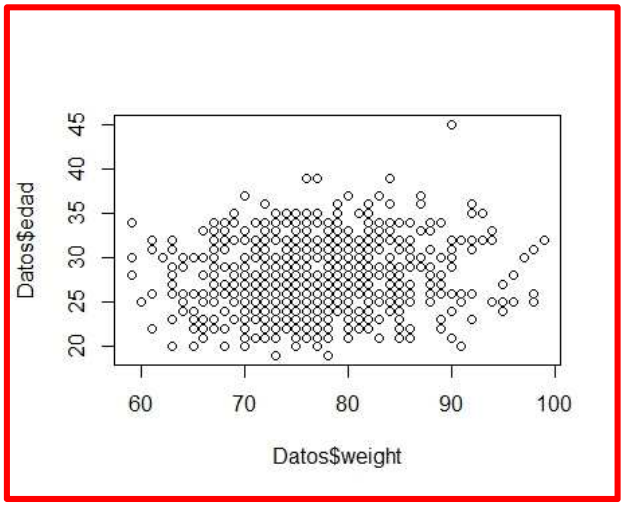
Respuesta

Podemos hacer varias cosas de las solicitadas. Vamos a plantear tres pruebas. Un test de correlación entre variables (vamos a hacer dos de ellas, aunque podríamos hacer una comparación entre un conjunto de variables). Vamos a realizar un test de hipótesis sobre si la edad media de los jugadores de España es la misma que la de los jugadores de Egipto. Y, por último, vamos a calcular la regresión ente el peso y la altura de los jugadores.

Correlación

Vamos a estudiar si existe relación entre la edad de los jugadores y el peso.

Lo primero que hacemos es obtener un gráfico de dispersión. Como podemos observar, la nube de puntos con forma esférica ya nos indica que entre estas variables no hay relación significativa.



El test de correlación nos aporta el siguiente resultado.

```
> cor.test(Datos$edad, Datos$weight) Pearson's product-moment correlation

data:  Datos$edad and Datos$weight

t = 4.6339, df = 734, p-value = 4.246e-06
```

alternative hypothesis: true correlation is not equal to 0, 95 percent confidence interval:

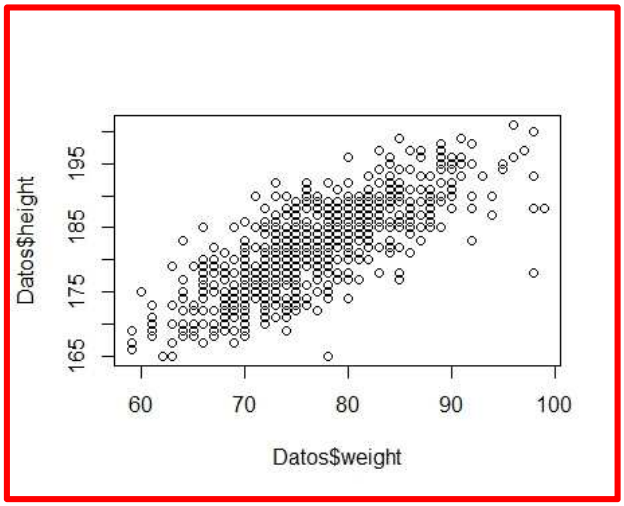
(0.09751237, 0.23795859)

sample estimates: cor = **0.168591**

Con un p-valor del estadístico de 0,000 rechazamos la hipótesis nula de que las variables son independientes (en realidad que la correlación es cero). El valor del coeficiente de 0,1686 nos indica que entre las variables no se puede establecer una relación de causa efecto.

Solo a nivel del ejercicio, vamos a ver un caso que sí tendría relación (peso y altura).

El grafico de dispersión queda: Se puede comprobar como el incremento de una variable se ve relacionado con el incremento de la otra.



En cuanto al test, obtenemos los siguientes resultados.

> cor.test(Datos$height, Datos$weight) Pearson's product-moment correlation

data: Datos$height and Datos$weight

t = 32.913, df = 734, p-value < **2.2e-16**

alternative hypothesis: true correlation is not equal to 0, 95 percent confidence interval:

(0.7411608, 0.7997204)

sample estimates: cor = **0.7720743**

En este caso, rechazamos que el valor de la correlación es cero y obtenemos el valor de 0,772.

Regresión

Los pasos que seguimos para obtener un modelo de región son los siguientes:

- Obtenemos el gráfico de dispersión (ya visto antes).
- Calculamos el valor del coeficiente (ya visto antes también).
- Calculamos el modelo lineal.
- Obtenemos los estadísticos del modelo.
- Obtenemos los coeficientes.
- Pintamos los datos con el modelo ajustado.

> cor.test(Datos$height, Datos$weight) Pearson's product-moment correlation

data: Datos$height and Datos$weight

t = 32.913, df = 734, p-value < **2.2e-16**

alternative hypothesis: true correlation is not equal to 0, 95 percent confidence interval: (0.7411608, 0.7997204)

```
sample estimates: cor = 0.7720743

> modelo_altura <- lm (Datos$height~Datos$weight)

> summary(modelo_altura)

Call: lm(formula = Datos$height ~ Datos$weight)

Residuals:          Min       1Q       Median    3Q       Max
                    -19.8027  -2.9667  0.1911    2.9923   12.6911
Coefficients:       Estimate             Std. Error t value  Pr(>|t|)
(Intercept)         125.30714            1.74247    71.91    <2e-16 ***
Datos$weight        0.73975              0.02248    32.91    <2e-16 ***
Residual standard error: 4.408 on 734 degrees of freedom

Multiple R-squared: 0.5961,    Adjusted R-squared: 0.5955

F-statistic: 1083 on 1 and 734 DF,  p-value: < 2.2e-16

> modelo_altura$coefficients:  (Intercept)          Datos$weight
                               125.3071433         0.7397501
```

A pesar de que el coeficiente de correlación era alto, superior a 0,75, el modelo de regresión con la forma Altura = 125,31 + 0,74 * peso, tiene un valor de R = 0,59. Desde mi punto de vista, bajo para realizar estimaciones con el mismo.

Test de hipótesis

Hay muchas formas de plantear esta técnica. En particular vamos a comparar si la edad media de los jugadores de España y Egipto son la misma. Partimos de la siguiente hipótesis nula.

$$H_o \equiv \mu_{España} = \mu_{Portugal}$$

Si los datos no contradicen la hipótesis nula, entonces nada se opone a aceptar que las edades medias de los jugadores de España y Egipto son iguales (tener en cuenta que Egipto tiene un jugador con 45 años).

Si queremos verlo de una forma sencilla, podemos calcular las medias de cada grupo y luego ver la diferencia entre ellas. Los datos que obtenemos en R son los siguientes:

Media de edad de los jugadores españoles = 28.3913

Media de edad de los jugadores egipcios = 28.91304

Si miramos las desviaciones estándar, para España es 3.639919 y para Egipto es 5.080382.

Si comparamos las medias de (España - Egipto) = -0.5217391.

La pregunta es si esta diferencia es estadísticamente significativa. Para comprobarlo, podemos utilizar el test de Wilcoxon.

```
> wilcox.test(x = España, y = Egipto, alternative = "two.sided", mu = 0, + paired = FALSE, conf.int = 0.95)

        Wilcoxon rank sum test with continuity correction

data:  España and Egipto

W = 263, p-value = 0.9824

alternative hypothesis: true location shift is not equal to 0, 95 percent confidence interval: (-2.999954, 2.000016)

sample estimates: difference in location = -1.703908e-05
```

Con un p-valor para el test de 0,9824, nada se opone a aceptar la hipótesis nula, y por lo tanto tenemos que concluir que la media de edad entre estas dos selecciones es la misma.

En el otro lado, podemos comparar Argentina con una media de edad de 29.17391 años, frente a Francia con una media de 25.91304, y desviaciones típicas de 3.725248 y 3.812814 respectivamente, nos lleva a una diferencia de (Argentina -Francia) = 3.26087. Parece que existirá diferencia entre estas selecciones. De hecho el test de Wilcoxon nos muestra un p-valor de 0,010, lo que nos obliga a rechazar la hipótesis nula y por lo tanto concluir que existe diferencia entre estos países.

```
> wilcox.test(x = Argentina, y = Francia, alternative = "two.sided", mu = 0, + paired = FALSE, conf.int = 0.95)

        Wilcoxon rank sum test with continuity correction

data:  Argentina and Francia

W = 381.5, p-value = 0.009972

alternative hypothesis: true location shift is not equal to 0, 95 percent confidence interval: (0.999969, 6.000018)

sample estimates: difference in location = 3.99997
```

Ejercicio 5

Representación de los resultados a partir de tablas y gráficas.

Respuesta

En el ejercicio se han ido mostrando los gráficos que se han considerado convenientes para documentar los pasos seguidos. Nos falta mostrar una tabla con los registros (los primeros) para ver como vienen en el fichero una vez importados a R. Podríamos haber insertado la propia hoja de Excel con los datos en formato CSV leídos correctamente.

RStudio Source Editor — Datos — Filter

	team	number	position	birth_date	shirt_name	club	height	weight	league	age	name	caps	edad
1	Argentina	1	GK	1986-02-10	GUZMÁN	Tigres UANL	192	90	MEX	32.33973	Nahuel Guzmán	6	32
2	Argentina	2	DF	1987-03-18	MERCADO	Sevilla FC	181	81	ESP	31.24110	Gabriel Mercado	20	31
3	Argentina	3	DF	1992-08-31	TAGLIAFICO	AFC Ajax	169	65	NED	25.78630	Nicolás Tagliafico	4	26
4	Argentina	4	DF	1986-09-20	ANSALDI	Torino FC	181	73	ITA	31.73151	Cristian Ansaldi	5	32
5	Argentina	5	MF	1986-01-30	BIGLIA	AC Milan	175	73	ITA	32.36986	Lucas Biglia	57	32
6	Argentina	6	DF	1987-03-17	FAZIO	AS Roma	199	85	ITA	31.24384	Federico Fazio	9	31
7	Argentina	7	MF	1988-06-29	BANEGA	Sevilla FC	175	73	ESP	29.95890	Éver Banega	62	30
8	Argentina	8	DF	1991-10-28	ACUÑA	Sporting CP	172	77	POR	26.62740	Marcos Acuña	10	27
9	Argentina	9	FW	1987-12-10	HIGUAÍN	Juventus FC	184	75	ITA	30.50959	Gonzalo Higuaín	71	31
10	Argentina	10	FW	1987-06-24	MESSI	FC Barcelona	170	72	ESP	30.97260	Lionel Messi	124	31
11	Argentina	11	MF	1988-02-14	DI MARÍA	Paris Saint-Germain FC	178	75	FRA	30.32877	Ángel Di María	94	30
12	Argentina	12	GK	1986-10-16	ARMANI	CA River Plate	189	85	ARG	31.66027	Franco Armani	0	32
13	Argentina	13	MF	1992-12-15	MEZA	CA Independiente	180	76	ARG	25.49589	Maximiliano Meza	2	25
14	Argentina	14	DF	1984-06-08	MASCHERANO	Hebei China Fortune FC	174	73	CHN	34.01644	Javier Mascherano	143	34
15	Argentina	15	MF	1993-02-15	LANZINI	West Ham United FC	167	66	ENG	25.32603	Enzo Pérez	23	25
16	Argentina	16	DF	1990-03-20	ROJO	Manchester United FC	189	82	ENG	28.23562	Marcos Rojo	56	28
17	Argentina	17	DF	1988-02-12	OTAMENDI	Manchester City FC	181	81	ENG	30.33425	Nicolás Otamendi	54	30
18	Argentina	18	DF	1990-07-13	SALVIO	SL Benfica	167	69	POR	27.92055	Eduardo Salvio	9	28
19	Argentina	19	FW	1988-06-02	AGÜERO	Manchester City FC	172	74	ENG	30.03288	Sergio Agüero	85	30
20	Argentina	20	MF	1996-04-09	LO CELSO	Paris Saint-Germain FC	177	75	FRA	22.18082	Giovani Lo Celso	5	22
21	Argentina	21	FW	1993-11-15	DYBALA	Juventus FC	177	73	ITA	24.57808	Paulo Dybala	12	25
22	Argentina	22	MF	1996-01-21	PAVÓN	CA Boca Juniors	169	65	ARG	22.39452	Cristian Pavón	5	22
23	Argentina	23	GK	1981-09-28	CABALLERO	Chelsea FC	186	80	ENG	36.70959	Willy Caballero	3	37
24	Australia	1	GK	1992-04-08	RYAN	Brighton & Hove Albion FC	184	82	ENG	26.18356	Mathew Ryan	44	26

Showing 1 to 26 of 736 entries, 13 total columns

Ejercicio 6

Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Respuesta

Los datos elegidos, igual no eran los más adecuados para el objetivo de la práctica, ya que la misma se fundamentaba en la lectura de los datos, análisis de los mismos y corrección de aquellos que se consideraran errores, atípicos, fuera de rango, o de cualquier otro tipo que

requiriera ciertas operaciones por parte del alumno, pero la selección se ha hecho sobre un conjunto de datos que no presentaba errores y que tenía una estructura perfecta para hacer análisis de todo tipo (sencillos) en los que se mostraran las técnicas aprendidas durante el curso con esta asignatura y otras del Máster.

He intentado tratar todos los temas solicitados, pero creo que el conjunto daba para haber intentado responder a otro tipo de preguntas, tal como se proponía al comienzo. Se me ocurre intentar contestar a preguntas del tipo:

- ¿La edad, corpulencia, experiencia influye en los resultados de un torneo de este tipo?
- ¿Son los equipos similares en edad, corpulencia, experiencia?
- ¿El equipo de origen de los jugadores condiciona los resultados obtenidos?
- ¿Es el fútbol de un continente el que obtiene mejores resultados?
- Otras muchas que se me pueden ocurrir.

En todo caso, sería interesante poner en común estos datos con otros de la competición, cómo goles a favor o en contra, posición en el torneo (campeón, finalista, semifinalista, cuartos, fase de grupos), minutos jugados, etc.

Ejercicio 7

Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Respuesta

El código se remite en el fichero “PRAC 2 Análisis de datos.R”

El fichero con los datos originales es “all_wc_18_players_fifa.csv”

No se ha generado un nuevo fichero al no tener que modificar los datos.

El fichero con los datos de la WIKI es “Enlace GitHub.txt”

CONTRIBUCIONES

Incluyo a mi compañero de la PRAC1 porque se merece estar en este trabajo.

Contribuciones	Firma
Ejercicio 1	LAP, JLMR
Ejercicio 2	LAP, JLMR
Ejercicio 3.1	LAP, JLMR
Ejercicio 3.2	LAP, JLMR
Ejercicio 4.1	LAP, JLMR
Ejercicio 4.2	LAP, JLMR
Ejercicio 4.3	LAP, JLMR
Ejercicio 5	LAP, JLMR

Ejercicio 6	**LAP, JLMR**
Ejercicio 7	**LAP, JLMR**

Recursos

Los siguientes recursos son de utilidad para la realización de la práctica:

- Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
- Megan Squire (2015). Clean Data. Packt Publishing Ltd.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.
- Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
- Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.
- Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.
- Tutorial de Github https://guides.github.com/activities/hello-world.

Bibliografía

Criterios de valoración

Todos los apartados son obligatorios. La ponderación de los ejercicios es la siguiente:

- Los apartados 1, 2 y 6 valen 0,5 puntos.
- Los apartados 3, 5 y 7 valen 2 puntos.
- El apartado 4 vale 2,5 puntos.

Se valorará la idoneidad de las respuestas, que deberán ser claras y completas. Las diferentes etapas deberán justificarse y acompañarse del código correspondiente. También se valorará la síntesis y claridad, a través del uso de comentarios, del código resultante, así como la calidad de los datos finales analizados. Documentación del programa del curso.