

PRACTICA 2: LIMPIEZA Y VALIDACIÓN DE LOS DATOS.

Carlos Lavado Mahia, Dionisio González Jiménez

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

El dataset elegido originalmente de kaggle, “*country_profile_variables.csv*”. consiste en la muestra de varios indicadores macroeconómicos para una gran variedad de países (PIB, tasa de crecimiento del PIB, PIB per cápita, agricultura, industria, servicios, empleados en varios sectores, desempleados, entre otros) a nivel global durante el año 2017.

En particular, queremos encontrar patrones mundiales para explicar cómo influyen diferentes factores sobre el Producto Interior Bruto de un país. Nuestra hipótesis inicial sería: ¿Puede modelarse la riqueza de un país mediante el estudio de variables macroeconómicas?

Por ejemplo, quisiéramos ver si es posible modelar cómo contribuye cada uno de los sectores económicos (agricultura, industria, servicios) al PIB per cápita.

Otro de nuestros objetivos es el de comprobar si es posible obtener correlaciones significativas entre la riqueza de un país y algunas variables demográficas.

Por último, nos gustaría estudiar si existen diferencias significativas en el crecimiento de la riqueza entre países desarrollados y países en vías de desarrollo.

2. Integración y selección de los datos de interés a analizar.

- **Integración:** como hemos dicho anteriormente, el dataset es originalmente de kaggle, creado el csv por la usuaria Anuradha Satyanarayana y que contiene 50 columnas. No hemos hecho fusión con otros dataset y además hemos considerado algunas columnas irrelevantes para nuestro estudio. Por ello, en el proceso de integración y selección, solo hemos seleccionado una serie de columnas que creemos que pueden ser relevantes para nuestros análisis.
- **Selección:** el proceso de selección ha consistido en analizar qué columnas del dataset eran interesantes para realizar posteriormente nuestros análisis, debido a que existían muchas columnas que no las hemos considerado relevantes para llevar a cabo nuestros estudios.

Las columnas elegidas en el proceso de selección han sido las siguientes:

- **Country**→ nombre del país.
- **Region**→ nombre de la región.
- **Population in thousands**→ población en miles de personas.
- **Population density**→ población por kilómetro cuadrado.
- **Population growth rate**→ tasa de crecimiento de la población (media anual).
- **Urban population (%)**→ porcentaje de población urbana
- **GDP**→ Producto Interior Bruto (PIB).
- **GDP growth rate**→ tasa de crecimiento del PIB.
- **GDP per capita**→ PIB per cápita.

- **Economy: Agriculture**→ porcentaje que aporta el sector agricultura al Valor Agregado Bruto (VAB).
- **Economy: Industry**→ porcentaje que aporta el sector industrial al VAB.
- **Economy: Services and others activities**→ porcentaje que aporta el sector servicios y otras actividades al VAB.
- **Employment: Agriculture**→ porcentaje de empleados en agricultura.
- **Employment: Economy**→ porcentaje de empleados en el sector industrial.
- **Employment: Services and others activities**→ porcentaje de empleados en el sector servicios y otras actividades.
- **Education: Government expenditure**→ gasto público en educación (medido en porcentaje).

Una vez seleccionadas las columnas, hemos modificado sus nombres, con tal de facilitar el trabajo con estos ya que eran demasiado largos.

Posteriormente, hemos transformado varias columnas de datos, ya que varias de las variables del dataset estaban en formato *carácter* (seguramente dado que existían valores NA no numéricos en sus columnas); por lo que han sido convertidas a formato numérico para facilitar sus análisis posteriores, como podemos ver en la siguiente figura (aprovechamos para adjuntar un breve resumen del dataset):

Ilustración 1. Transformación de columnas y resumen del dataset empleado en el proyecto

```
# Convertimos las columnas que no se han interpretado con el tipo deseada.
country_profiles$Region<-as.factor(country_profiles$Region)
country_profiles$pop_growth<-as.numeric(country_profiles$pop_growth)
country_profiles$gdp_growth<-as.numeric(country_profiles$gdp_growth)
country_profiles$eco_agri<-as.numeric(country_profiles$eco_agri)
country_profiles$empl_agri<-as.numeric(country_profiles$empl_agri)
country_profiles$empl_industry<-as.numeric(country_profiles$empl_industry)
country_profiles$empl_services<-as.numeric(country_profiles$empl_services)
country_profiles$educ_exp<-as.numeric(country_profiles$educ_exp)
```

```
##      country      Region      population      pop_dens
## Length:229      Length:229      Min. : 1      Min. : 0.1
## Class :character Class :character 1st Qu.: 431 1st Qu.: 35.9
## Mode :character Mode :character Median : 5448 Median : 88.1
## Mean : 32757 Mean : 462.8
## 3rd Qu.: 19193 3rd Qu.: 222.8
## Max. :1409517 Max. :25969.8
##      gdp      gdp_growth      gdp_xcap      eco_agri
## Min. : -99      Length:229      Min. : -99      Length:229
## 1st Qu.: 2078      Class :character 1st Qu.: 1208      Class :character
## Median : 16251      Mode :character Median : 4836      Mode :character
## Mean : 321434      Mean : 14252
## 3rd Qu.: 117955      3rd Qu.: 16344
## Max. :18036648      Max. :169492
##      eco_industry      eco_services      empl_agri      empl_industry
## Min. : -99.00      Min. : -99.00      Length:229      Length:229
## 1st Qu.: 15.40      1st Qu.: 47.30      Class :character Class :character
## Median : 25.50      Median : 59.50      Mode :character Mode :character
## Mean : 15.96      Mean : 46.41
## 3rd Qu.: 32.80      3rd Qu.: 70.70
## Max. : 79.90      Max. : 94.00
##      empl_services      pop_growth      urban_pop      educ_exp
## Length:229      Length:229      Min. : 0.00      Length:229
## Class :character Class :character 1st Qu.: 39.80      Class :character
## Mode :character Mode :character Median : 59.90      Mode :character
## Mean : 59.51
## 3rd Qu.: 79.60
## Max. :100.00
```

3. Limpieza de los datos.

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

La detección de 0s, se ha hecho a través de R con la cláusula *which*. Tras una inspección exhaustiva, tan solo se han detectado 0s en la columna de “urban_pop”, en concreto para los países “Wallis and Futuna Islands” y “Tokelau”. Sabemos que es factible que en estos territorios no haya población urbana, por lo que no se ha considerado necesario su tratamiento.

Ilustración 2. Ejemplo de búsqueda de 0s (población urbana)

```
# Analizamos si tenemos valores cero en población urbana
which(country_profiles$urban_pop == 0.0)
```

```
## [1] 205 225
```

Por otro lado, para comprobar si tenemos valores vacíos en cada una de las variables o columnas del dataset, se utiliza “*which(is.na(country_profiles\$columna))*”. Aquí hemos observado varias columnas que presentaban valores vacíos; en muchos casos se trataba de datos que, en el dataset original, presentaban un valor “...”; y al convertir la columna a numérico, estos se transformaron en NA por coerción.

Ilustración 3. Ejemplo de búsqueda de valores vacíos (empleo agricultura)

```
which(is.na(country_profiles$empl_agri))
```

```
## [1] 4 7 8 30 60 70 101 109 127 150 151 155 170 171 176 211
```

Estos valores vacíos, han sido transformados mediante la librería *missForest*. Esto se explicará más detalladamente en el siguiente punto, puesto que convertiremos algunos valores extremos a NA antes de proceder a la imputación de valores.

3.2. Identificación y tratamiento de valores extremos.

Los valores extremos han sido identificados a través de la función *box-plot* incorporada en R; en concreto a través del parámetro *\$out*, el cual devuelve los campos que se sitúan a más de 1,5 de distancia del rango intercuartílico.

Ilustración 4. Ejemplo de búsqueda de valores extremos (población)

```
boxplot.stats(country_profiles$population)$out
```

```
## [1] 164670 209288 1409517 49066 81340 97553 104957 64980 82114
## [10] 1339180 263991 81163 59360 127484 49700 129163 53371 190886
## [19] 197016 104918 50982 143990 56717 69038 80745 66182 57310
## [28] 324460 95541
```

Sin embargo, lo que hemos observado mayoritariamente, ha sido que **los valores extremos eran razonables**, por lo que no hemos considerado necesaria su modificación. Por ejemplo, el archipiélago Tokelau presentaba un valor del 0% de población urbana, mientras que a Macao (China) se le asignaba uno del 100%. Tras unas comprobaciones manuales a través de nuestros navegadores, vimos que estos son datos oficiales, a pesar de considerarse obviamente valores extremos por la función *box-plot*.

- **Population:** observamos que determina algunos valores extremos, pero dentro de lo normal, ya que a pesar de ser valores altos, son razonables a la variable de la que se trata. Podemos destacar China con 1.409.517 miles de personas o la India con 1.339.180 miles de personas, pero se trata de datos reales y no considerados como valores extremos.

```
## [1] 164670 209288 1409517 49066 81340 97553 104957 64980 82114
## [10] 1339180 263991 81163 59360 127484 49700 129163 53371 190886
## [19] 197016 104918 50982 143990 56717 69038 80745 66182 57310
## [28] 324460 95541
```

- **Pop_dens:** con respecto a la densidad de población, también obtenemos valores extremos con boxplot; pero valores coherente en relación a la variable que estamos analizando, ya que podemos ver el ejemplo claro de 20.821,6 de densidad de población que destaca para China.

```
## [1] 584.8 1963.9 1265.0 664.5 1227.0 870.1 7014.2 20821.6 3457.1
## [10] 1800.0 594.6 1454.4 1346.4 623.2 674.8 25969.8 568.0 505.2
## [19] 524.3 556.7 8155.5 1180.0 817.4
```

- **Pop_growth:** con respecto al crecimiento de la población, se observan datos extremos pero se trata de valores coherentes.

```
## [1] 5.4 6.0 6.5 6.6
```

- **Urban_pop:** con respecto a la tasa de crecimiento de la población urbana no detectamos valores extremos utilizando `boxplot.stats$out`.
- **Gdp:** el PIB devuelve valores extremos a través de la ejecución de `boxplot.stats$out`, pero observamos que se tratan de datos coherentes; ya que son varios los países que tienen una producción muy alta, como puede destacar China con 11.158.457 millones de dólares, pues es una de las grandes potencias mundiales.

```
## [1] 632343 1230859 376967 455107 1772591 1552808 309236 11158457
## [9] 292080 301308 315917 2418946 3363600 2116239 861934 398563
## [17] 299413 1821580 4383076 296284 1140724 750318 494583 386578
## [25] 292449 477066 1377873 1326016 653219 292734 314571 1192955
## [33] 495694 670790 395168 717888 370296 2858003 18036648 344331
```

- **Gdp_growth:** en esta variable sí que observamos valores extremos, algunos considerados normales debido a que se trata únicamente de tasas de crecimiento negativas; y otros valores de -99, que son valores perdidos según lo observado en el dataset (se explicará más adelante). Se puede ver en la siguiente figura una tasa de -28,1 para Yemen, un país bastante pobre y que es coherente que haya tenido una tasa de crecimiento del PIB negativa. En cambio, los valores -99 sí que serán sustituidos a través de `missForest`.

```
## [1] -99.0 -99.0 -99.0 -20.3 -7.4 9.6 -99.0 -99.0 -99.0 -99.0 -99.0
## [13] -99.0 26.3 -99.0 -10.2 -99.0 -99.0 18.7 -99.0 -99.0 -99.0 -20.3
## [25] -5.3 -99.0 -9.9 -99.0 -6.2 -99.0 -99.0 -28.1
```

- **Gdp_xcap:** al igual que la variable PIB, el PIB per cápita nos muestra valores extremos, pero se trata de valores coherentes (como por ejemplo Liechtenstein).

```
## [1] 39896.4 51352.2 44117.7 40277.8 94399.9 43205.6 62132.0 42431.0
## [9] 78586.4 53149.3 42148.1 41686.2 50936.0 60513.6 169491.8 100160.8
## [17] 165870.6 44332.1 74185.5 73653.4 49240.2 52239.0 50687.5 80831.1
## [25] 40438.8 44162.4 56053.8
```

- **Eco_agri:** observamos que para la variable del sector agricultura existen valores -99, considerados valores perdidos que regularizaremos con missForest. También aparecen algunos porcentajes altos, pero se considera dentro los valores normales para esta variable; como por ejemplo Liberia, con un 70,8, pues se trata de un país de África para el que consideramos que es probable este crecimiento del sector agrícola.

```
## [1] -99.0 -99.0 -99.0 -99.0 40.5 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 45.0
## [13] -99.0 -99.0 70.8 39.9 -99.0 -99.0 -99.0 39.6 -99.0 -99.0 -99.0
## [25] 51.4 60.2 45.7 -99.0 -99.0 -99.0 -99.0
```

- **Eco_industry:** la variable sector industria también muestra bastantes valores extremos, si bien en su mayoría vuelve a tratarse de valores coherentes. El resto (valores -99) serán regularizados con missForest.

```
## [1] -99.0 -99.0 60.2 -99.0 70.0 73.1 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0
## [13] -99.0 -99.0 67.1 -99.0 -99.0 59.9 -99.0 -99.0 -99.0 -99.0 79.9 -99.0
## [25] -99.0 -99.0 -99.0
```

- **Eco_services:** con respecto al sector servicios, los valores que son detectados como extremos, son todos valores -99 que sí que serán regularizados.

```
## [1] -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99
## [20] -99 -99
```

- **Empl_agri:** el empleo en agricultura nos muestra algún porcentaje alto (91,1) en Burundi, un país perteneciente a África y que podemos considerar razonable por lo que no será modificado. El resto de valores extremos (-99) serán regularizados con missForest para facilitar los análisis.

```
## [1] -99.0 -99.0 80.0 91.1 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0
## [13] -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0
```

- **Empl_industry:** con respecto a la empleabilidad en el sector industria, muestra varios valores extremos (-99) que son corregidos con el uso de missForest, así como algún valor concreto que entra dentro de los intervalos normales.

```
## [1] -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0
## [13] 54.1 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0
```

- **Empl_services:** con respecto a la variable empleo en servicios. todos los valores que se detectan como extremos deben ser regularizados, al tratarse de valores perdidos.

```
## [1] -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99 -99
```

- **Educ_exp:** por último, la variable del gasto público en educación detecta varios valores perdidos (-99), que también serán regularizados con missForest para facilitar sus estudios posteriores.

```
## [1] -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0
## [13] -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0
## [25] -99.0 12.5 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0
## [37] -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0 -99.0
```

Tras una revisión de las columnas de datos seleccionados, se ha descubierto que **muchas disponían del valor -99 en descontadas ocasiones**. Dada la repetición de este suceso, se ha llegado a la conclusión de que estos campos han sido marcados por la usuaria original de Kaggle al tratarse de valores perdidos. Por tanto, han sido pasados a valores NA.

Una vez cambiados los valores -99 a NA, se ha procedido a la imputación de valores perdidos. Para ello, nos hemos apoyado del algoritmo basado en aprendizaje automático **missForest**, útil para la imputación en conjuntos de datos mixtos multidimensionales.

```
# Aplicamos algoritmo a los datos para limpiarlos.
country_profiles_clean <- missForest(country_profiles)
```

```
## missForest iteration 1 in progress...done!
## missForest iteration 2 in progress...done!
## missForest iteration 3 in progress...done!
## missForest iteration 4 in progress...done!
## missForest iteration 5 in progress...done!
```

```
country_profiles <- country_profiles_clean$ximp
```

MissForest es un algoritmo de imputación de datos perdidos, que imputa inicialmente los datos mediante la media o la moda; después, para cada variable en la que haya valores perdidos, MissForest elabora un *random forest* en la parte observada, con tal de predecir la parte faltante.

4. Análisis de los datos.

NOTA INTRODUCTORIA: El nivel de confianza para todos los análisis que lo requieran, será 5%.

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

En primer lugar, queremos realizar una **regresión lineal múltiple**, para observar cómo influye cada uno de los tres sectores económicos (agricultura, industria y servicios) sobre el PIB por cápita. Para ello, seleccionaremos tanto la columna del PIB por cápita, como las seis columnas específicas relacionadas con los sectores económicos (las 3 con el peso sobre el VAB de cada sector, y las 3 con el porcentaje de empleo de cada sector).

En segundo lugar, estudiaremos la correlación entre varias columnas del dataset a través de los **coeficientes de Spearman** (en particular, emplearemos las columnas de PIB y el resto de columnas no empleadas en la regresión, es decir, las variables demográficas).

Por último, realizaremos un contraste de hipótesis mediante un **test de Mann-Whitney**, para comprobar si existen diferencias estadísticamente significativas entre el crecimiento del PIB de Asia y Europa (lo que representaría a Oriente vs Occidente, o Países en vías de desarrollo vs Países desarrollados).

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Comprobación de normalidad

Para poder estudiar la normalidad de las columnas del dataset, se ha utilizado el **test Shapiro-Wilk** con un valor de confianza del 95%. A través de este, observamos que la mayoría de las variables no siguen una distribución normal, pues en todos los casos obtenemos un p-value muy bajo y tenemos por tanto que rechazar la hipótesis nula (normalidad).

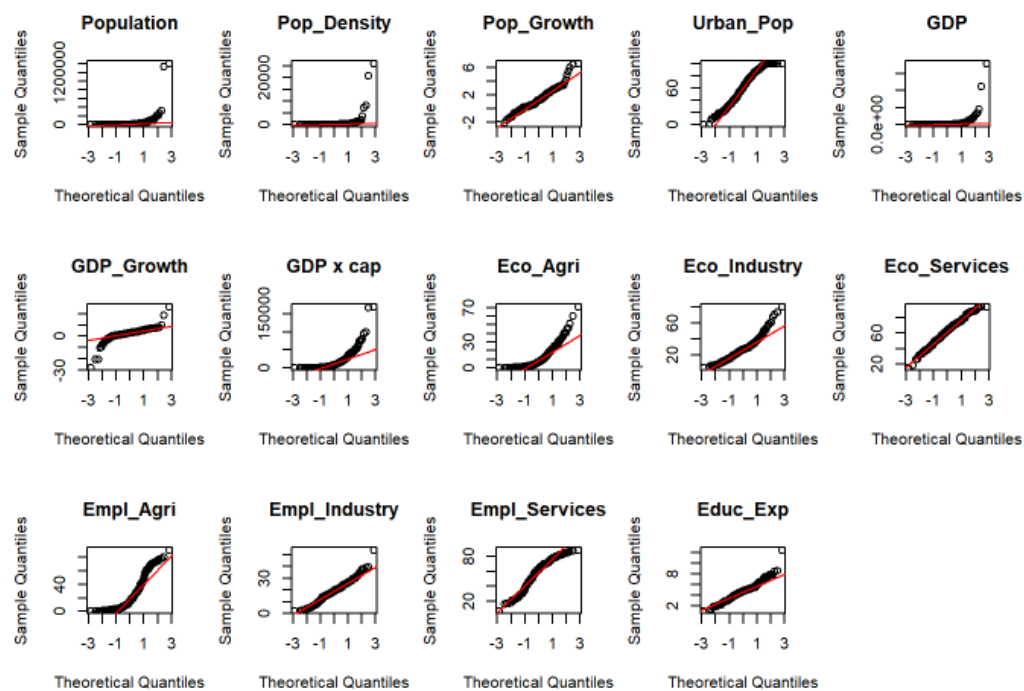
Además, se ha utilizado también el test de **Kolmogorov-Smirnov** para realizar pruebas adicionales sobre el estudio de la normalidad. Sin embargo, las conclusiones no han variado significativamente.

Por ello, hemos procedido a realizar **QQPlots**, con tal de investigar si la mayoría de puntos de cada columna se alinean con la línea que representa la distribución normal.

Tras esta inspección visual, hemos decidido que **no es posible asumir normalidad en los datos**, y por tanto esto delimitará los métodos no paramétricos de análisis que emplearemos en el proyecto.

En la página siguiente, se adjunta el output con los **QQPlots** relativos a cada una de las columnas del dataset.

Ilustración 5. QQPlots de todas las columnas del dataset

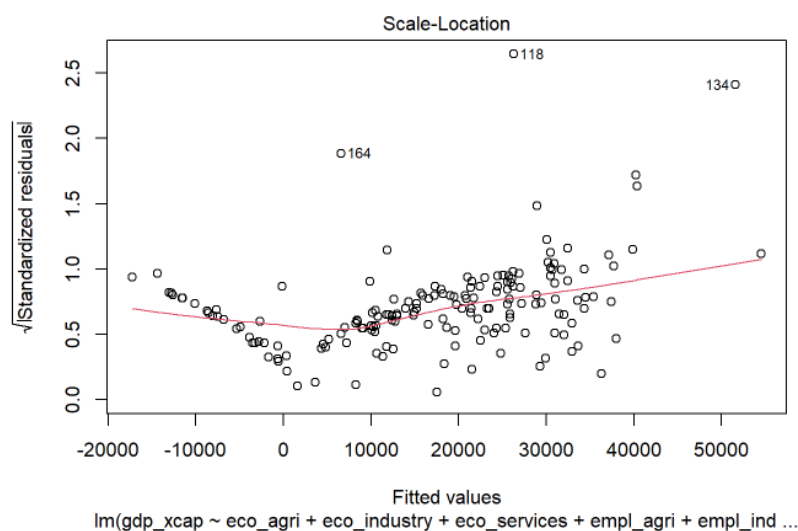


Comprobación de homocedasticidad (homogeneidad de las varianzas)

Las asunciones de homocedasticidad son distintas en cada uno de los análisis que realizaremos.

En primer lugar, la homocedasticidad de la **regresión lineal múltiple** debe ser investigada a través de los gráficos de los residuos, ya que estos deben ser homocedásticos. Esto se ha realizado al final del modelo, una vez obtenidos los residuos. A continuación, **se adjunta el gráfico de varianza de los residuos:**

Ilustración 6. Comprobación de homocedasticidad - Regresión lineal múltiple



Como podemos apreciar, se observa un **cierto patrón de heterocedasticidad** en los datos, cuya varianza va aumentando a partir de 0. Esto significa que, si bien los parámetros de nuestro modelo son correctos, los p-values podrían ser inferiores a los valores que cabría esperar. Por tanto, las conclusiones de nuestro modelo pueden estar sujetas a la sobreestimación.

Por otro lado, el **coeficiente de correlación de Spearman no asume homocedasticidad en los datos de entrada**.

Por último, para el **contraste de hipótesis mediante Mann-Whitney**, con tal de comprobar si existen diferencias estadísticas significativas entre el crecimiento de PIB en Europa y Asia, sí ha sido necesario comprobar que la varianza de la variable *gdp_growth* para todas las regiones sea homocedástica. A continuación, se presenta el análisis de Fligner-Killeen realizado:

Ilustración 7. Comprobación de homocedasticidad - Mann-Whitney

```
# Comprobamos igualdad de varianzas.
fligner.test(gdp_growth ~ Region, data=country_profiles)

##
## Fligner-Killeen test of homogeneity of variances
##
## data:  gdp_growth by Region
## Fligner-Killeen:med chi-squared = 12.924, df = 21, p-value = 0.9113
```

El p-value es muy alto, por lo tanto **no podemos rechazar la hipótesis nula que establece que la varianza de crecimiento del PIB en todas las regiones es similar**.

4.3. 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.

Primer Análisis - Regresión Lineal Múltiple

Nuestro primer objetivo, es el de descubrir el impacto que tienen varias variables en el PIB por cápita de una economía. En particular, queremos comprobar **cómo contribuye cada uno de los tres sectores económicos principales (agricultura, industria, servicios) a esta variable**. Para ello, realizaremos una regresión lineal múltiple.

Si bien nuestros datos no siguen una distribución normal, es posible la realización de regresiones lineales, ya que estas no asumen normalidad en los datos. En todo caso, deberemos comprobar que los residuos del modelo creado sigan una distribución normal (lo que será realizado al final del presente análisis).

Para poder realizar el presente análisis, haremos uso de la validación cruzada. En particular, dividiremos los datos a través de métodos de estratificación con la función `holdout()`, incorporada en la librería `rminer`. Con tal de asegurar una estratificación controlada, aplicaremos el parámetro “order”, ya que de otro modo, dada la cantidad de datos en la muestra, obtendríamos resultados muy distintos en cada ejecución si los estratos se crearan de forma aleatoria.

Ilustración 8. División de los datos en entrenamiento y validación

```
# Dividimos datos en entrenamiento y validación, con un ratio de 3/4.  
# Mantendremos los datos estratificados por región.  
h<-holdout(country_profiles$Region, ratio = 3/4, mode="order")  
data_train<-country_profiles[h$str,]  
data_test<-country_profiles[h$ts,]
```

En primer lugar, a través de la función `lm()`, entrenamos al dataset de entrenamiento para obtener el modelo de regresión lineal múltiple.

Ilustración 9. Análisis de regresión lineal múltiple

```
# Creamos el modelo, y realizamos un resumen del mismo.  
regressor <- lm(gdp_xcap ~ eco_agri + eco_industry + eco_services + empl_agri + empl_industry + empl_services, data = data_train)  
summary(regressor)
```

```
##  
## Call:  
## lm(formula = gdp_xcap ~ eco_agri + eco_industry + eco_services +  
##     empl_agri + empl_industry + empl_services, data = data_train)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -27493 -11488  -4822    6650 143197   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)  -1.003e+06  3.272e+05  -3.064  0.00255 **   
## eco_agri       9.983e+03  3.508e+03   2.846  0.00499 **   
## eco_industry   1.034e+04  3.515e+03   2.942  0.00373 **   
## eco_services   1.044e+04  3.510e+03   2.973  0.00339 **   
## empl_agri      -4.321e+02  8.333e+02  -0.519  0.60475   
## empl_industry  -6.137e+02  8.433e+02  -0.728  0.46781   
## empl_services   8.284e+01  8.459e+02   0.098  0.92211   
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 20670 on 165 degrees of freedom  
## Multiple R-squared:  0.3256, Adjusted R-squared:  0.301   
## F-statistic: 13.27 on 6 and 165 DF, p-value: 3.033e-12
```

A través de los resultados del modelo de regresión, observamos que **el peso económico de los tres sectores principales sobre el VAB tiene cierta influencia en la explicación del PIB por cápita de un país** (p-value muy inferior al valor de significancia, 5%). Por el contrario, la proporción de empleo en cada sector no contribuye significativamente.

Una vez entrenado este modelo, podemos proceder a aplicar una predicción sobre la submuestra de validación, estudiando la correlación entre los valores reales y los predichos. De este modo, si obtenemos una alta correlación, significará que los datos de PIB per cápita han sido imputados con certeza por nuestro modelo de regresión.

Ilustración 10. Validación del modelo de regresión lineal múltiple

```
# Aplicamos el modelo sobre los datos de test, y estudiamos correlación entre datos reales y predichos.  
test_predict<-predict(regressor, newdata=data_test)  
cor(test_predict, data_test$gdp_xcap)
```

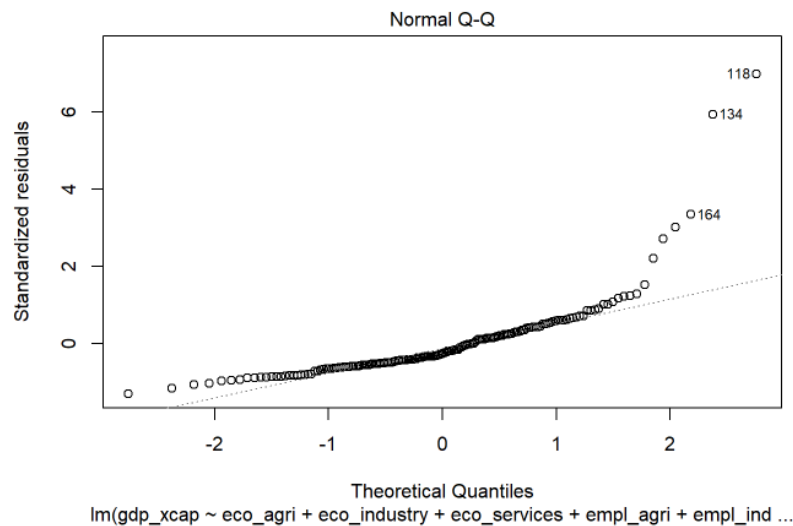
```
## [1] 0.5918964
```

La correlación entre los valores reales de la submuestra de validación, y los valores predichos tras aplicar sobre estos el modelo de regresión, es de aproximadamente un 60%. Por tanto, **podemos constatar que nuestro modelo presenta cierta eficacia ante nuevos datos.**

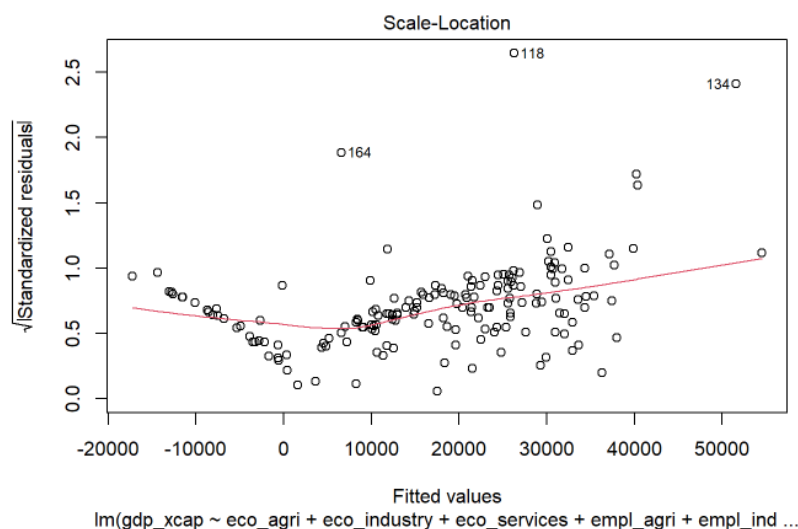
Ahora que ya hemos elaborado la regresión lineal y comprobado su eficacia mediante validación cruzada, comprobemos las suposiciones del modelo mediante los gráficos de los residuos (normalidad y homocedasticidad de los residuos).

Ilustración 11. Comprobación de las asunciones del modelo de regresión lineal múltiple

```
# Comprobamos normalidad a través del gráfico de residuos.
plot(regressor, which=2)
```



```
# Comprobamos homocedasticidad de los residuos.
plot(regressor, which=3)
```



En primer lugar, podemos observar que los puntos del primer gráfico se ajustan a la línea de cuantiles teóricos de la distribución normal. Por tanto, podemos confirmar que **nuestro modelo cumple con el supuesto de normalidad de los residuos**.

Por otro lado, **se observa una leve heterocedasticidad** en los datos del segundo gráfico, cuya varianza incrementa gradualmente a partir de 0. Por ello, si bien los parámetros son correctos, nuestros p-valores podrían haber sobreestimado la significancia de las covariables, ya que esta se ve afectada diferentemente por diversos intervalos de input.

Por tanto, podemos concluir que **el peso sobre el VAB de los sectores económicos podría contribuir de algún modo al PIB per cápita de una economía**. Sin embargo, dada la heterocedasticidad de los residuos, no deberíamos tomar esta conclusión a la ligera.

En el ejercicio siguiente, representaremos los estimadores del modelo para acabar de ver cómo evoluciona el PIB per cápita con cada covariable, y así tomar una decisión final.

Segundo Análisis - Coeficiente de correlación de Spearman

Con relación al **resto de variables demográficas** seleccionadas (como la densidad de población o el porcentaje de población urbana en un país), se ha decidido **estudiar su correlación con las figuras de PIB a través del método Spearman**, una alternativa no paramétrica empleada en casos en los que los datos no siguen una distribución normal. Como ya se ha explicado anteriormente, esta metodología tampoco asume homocedasticidad.

De este modo, a través de R, podemos observar las siguientes correlaciones de Spearman:

Ilustración 12. Matriz de correlaciones de Spearman

```
# Almacenamos las correlaciones de Spearman en una matriz.
corr.res <- cor(country_profiles[, c("gdp", "gdp_xcap", "gdp_growth", "urban_pop", "pop_dens", "educ_exp", "population", "pop_growth")], method="spearman")
corr.res
```

```
##          gdp  gdp_xcap gdp_growth urban_pop pop_dens
## gdp      1.000000000  0.3848117 -0.03725942  0.38252528  0.008620233
## gdp_xcap  0.384811684  1.00000000 -0.26810198  0.66556174  0.201000423
## gdp_growth -0.037259418 -0.2681020  1.00000000 -0.26111105  0.015524179
## urban_pop  0.382525283  0.6655617 -0.26111105  1.00000000  0.096537005
## pop_dens   0.008620233  0.2010004  0.01552418  0.09653700  1.000000000
## educ_exp   0.011627608  0.1458641 -0.10369403  0.06996132 -0.097466079
## population 0.552621606 -0.3469278  0.08247239 -0.11629183 -0.113163291
## pop_growth -0.182845130 -0.4287091  0.19406182 -0.22386803 -0.137773326
##          educ_exp population pop_growth
## gdp      0.01162761  0.55262161 -0.1828451
## gdp_xcap  0.14586405 -0.34692778 -0.4287091
## gdp_growth -0.10369403  0.08247239  0.1940618
## urban_pop  0.06996132 -0.11629183 -0.2238680
## pop_dens  -0.09746608 -0.11316329 -0.1377733
## educ_exp   1.00000000 -0.09317277 -0.1374265
## population -0.09317277  1.00000000  0.2029031
## pop_growth -0.13742646  0.20290314  1.0000000
```

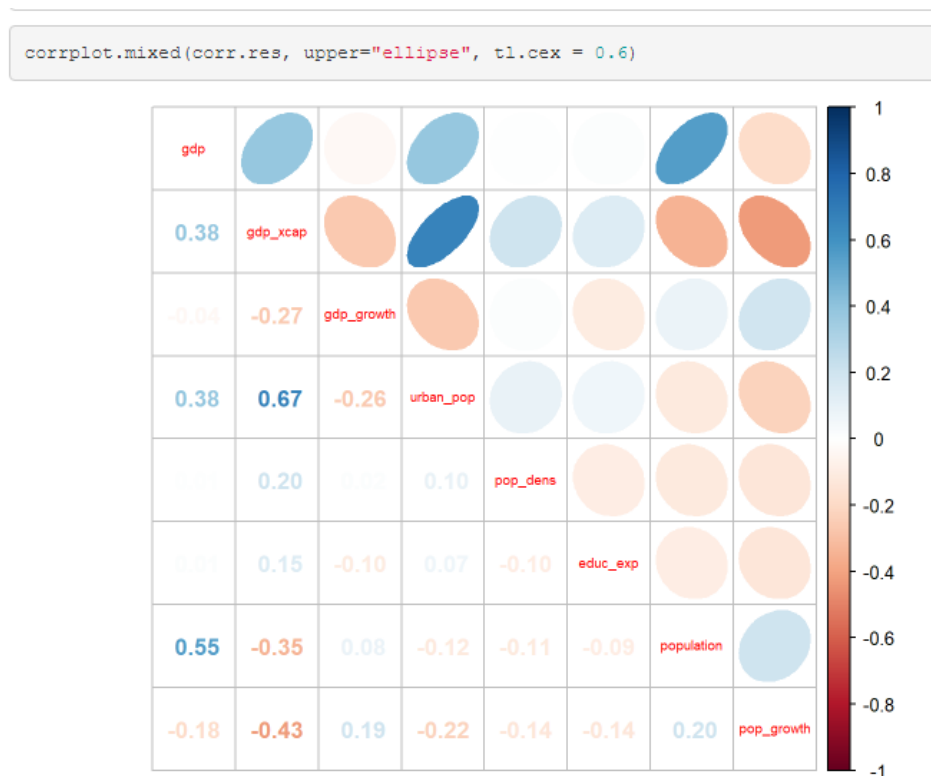
Los valores más cercanos a 1 en valor absoluto, son aquellos para los que existirá una correlación más fuerte entre las variables. En la anterior matriz, podemos observar una **alta correlación entre el porcentaje de población urbana y el PIB por cápita de un país (67%)**. Por otro lado, también hay una alta correlación entre el PIB y el número de habitantes (57%), si bien en este caso se trata de una obviedad.

Observamos por otro lado, curiosamente, como **el gasto en educación de un país no tiene correlación, por sí solo, con sus figuras de PIB**. Es decir, una mayor inversión en educación no parece traducirse necesariamente en un incremento del PIB, según los patrones de los países en 2017.

Además, la densidad de población de un país tampoco parece explicar sus figuras de PIB. Sin importar que un país esté densamente poblado o no, este aún puede registrar grandes figuras de PIB a través de otros factores más significativos.

A continuación, adjuntamos el gráfico de correlaciones de Spearman entre todas estas variables, en el que se puede observar como las correlaciones más significativas se presentan en forma de elipse:

Ilustración 13. Gráfico de correlaciones de Spearman



Tercer Análisis - Contraste de Hipótesis mediante el test de Mann-Whitney

Por último, para observar si podemos encontrar alguna **diferencia significativa entre los países de Oriente (generalmente, en vías de desarrollo) y de Occidente (generalmente desarrollados)**, se ha decidido realizar un test de Mann-Whitney para comprobar si existen diferencias estadísticas significativas entre el crecimiento de PIB anual de los países europeos y el de los países asiáticos.

El **test de Mann-Whitney** es un test no paramétrico, que por tanto no asume normalidad en los datos de entrada. Por otro lado, si asume homocedasticidad entre los conjuntos de datos empleados. Como se ha mostrado anteriormente a través del test de Fligner-Killeen sobre el crecimiento del PIB en las distintas regiones, nuestro modelo cumple con esta asunción.

El primer paso, ha sido el de investigar la dependencia entre los datos de Europa y los datos de Asia. Esto se debe, a que en función de si las variables son independientes o no, debería emplearse el test de Wilcoxon o el test de Mann-Whitney.

A continuación, se adjunta el **test de independencia** mediante Chi-Square llevado a cabo:

Ilustración 14. Comprobaciones de independencia para aplicar Wilcoxon o Mann-Whitney

```
# Primero, aplicamos un test Chi cuadrado para saber si las dos variables son dependientes entre ellas.  
# Dependiendo del resultado, emplearemos Wilcoxon o Mann-Whitney.  
chisq.test(table(country_profiles_europe, country_profiles_asia))
```

```
## Warning in chisq.test(table(country_profiles_europe, country_profiles_asia)):  
## Chi-squared approximation may be incorrect
```

```
##  
## Pearson's Chi-squared test  
##  
## data: table(country_profiles_europe, country_profiles_asia)  
## X-squared = 240, df = 225, p-value = 0.2348
```

Como obtenemos un p-value muy significativo (0.2348), no podemos rechazar la hipótesis nula del test de independencia de Chi-Square, que indica que no existe dependencia entre las variables. Es por ello que hemos procedido con la alternativa de Mann-Whitney.

El contraste de hipótesis mediante Mann-Whitney que hemos aplicado ha sido el siguiente:

$H_0 =$ No hay diferencia de medianas entre gdp_{growth} de Asia y de Europa

$H_1 =$ Sí hay diferencia de medianas entre gdp_{growth} de Asia y de Europa

Para llevar a cabo este análisis en R, lo hacemos a través de la función `wilcox.test`. Para indicar que queremos emplear la alternativa de Mann-Whitney, tenemos que establecer el parámetro “paired” a FALSE. A continuación, se presentan los resultados del análisis:

Ilustración 15. Aplicación del contraste de hipótesis mediante Mann-Whitney

```
wilcox.test(country_profiles_europe$gdp_growth, country_profiles_asia$gdp_growth, paired=FALSE)
```

```
##  
## Wilcoxon rank sum test with continuity correction  
##  
## data: country_profiles_europe$gdp_growth and country_profiles_asia$gdp_growth  
## W = 926, p-value = 0.0519  
## alternative hypothesis: true location shift is not equal to 0
```

Como podemos observar, **se obtiene un p-value ligeramente superior al 5%**, por lo que tendríamos que quedarnos con la hipótesis nula en este test a dos colas; y **descartar que exista una diferencia significativa de medianas entre las regiones**. Sin embargo, se trata de un p-value muy bajo, y la eficacia de tests como el de Mann-Whitney no tiene por qué ser perfecta. Es necesario acompañar esta conclusión de una representación visual.

En el siguiente apartado (tablas y gráficos), estudiaremos los box-plots de cada una de estas variables, para comprobar si, en efecto, una de ellas posee una mediana superior.

5. Representación de los resultados a partir de tablas y gráficas.

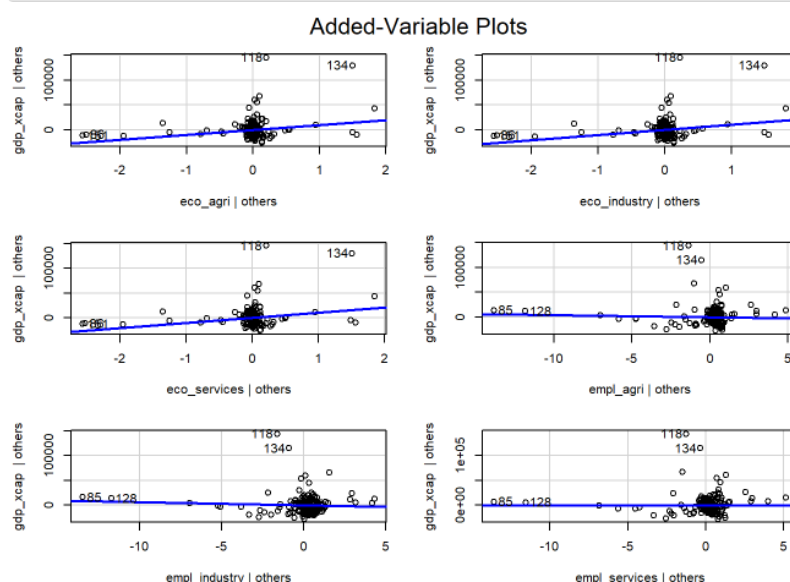
Gráficos y tablas de la regresión lineal múltiple

En el ejercicio 4, ya estudiamos los gráficos de los residuos para comprobar los supuestos del modelo, por lo que estos no serán aquí reinsertados.

En su lugar, hemos decidido representar los **Added-Variable Plots** de las distintas covariables del modelo, a través de la librería “car”. Estos nos permiten observar los patrones que pueden darse entre la variable independiente (gdp_xcap) y las variables dependientes. En principio, cabría esperar que las tres variables de peso económico de VAB presentaran un cierto patrón, si bien como hemos visto en la heterocedasticidad de los residuos, nuestros p-valores que indicaban significancia para estas tres columnas podrían estar sobreestimados.

Ilustración 16. Gráficos y tablas - Regresión lineal múltiple

```
avPlots(regressor)
```



Como podemos comprobar, el valor de gdp_xcap sube muy levemente a la vez que aumentan los valores de aporte al VAB de agricultura, industria y servicios; lo que no se traduce cuando aumenta la proporción de población empleada en cada uno de estos sectores.

Sin embargo, la curva de crecimiento es muy poco empinada, lo que nos lleva a reflexionar en que, efectivamente, nuestros p-valores de significancia de las covariables podrían haber sobreestimado y, por tanto, en algún caso puede que no se trate de una correlación significativa.

Gráficos y tablas del test de correlación de Spearman

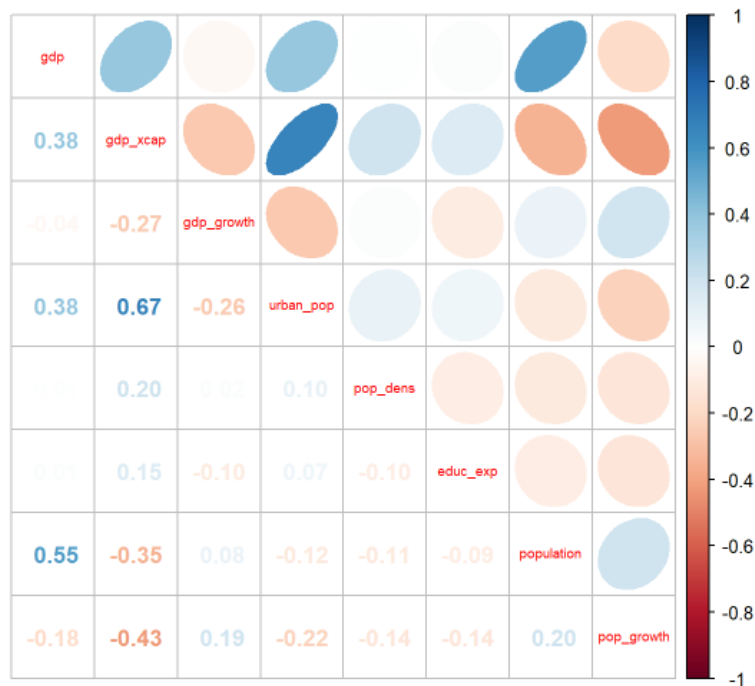
La tabla y el gráfico de correlaciones ya fueron adjuntados en la explicación del análisis del punto 4. Los volvemos a adjuntar para mayor claridad.

Ilustración 17. Gráficos y tablas - Correlaciones de Spearman

```
# Almacenamos las correlaciones de Spearman en una matriz.  
corr.res <- cor(country_profiles[, c("gdp", "gdp_xcap", "gdp_growth", "urban_pop", "pop_dens", "educ_exp", "population", "pop_growth")], method="spearman")  
corr.res
```

```
##           gdp  gdp_xcap gdp_growth urban_pop pop_dens  
## gdp      1.000000000  0.3848117 -0.03725942  0.38252528  0.008620233  
## gdp_xcap  0.384811684  1.00000000 -0.26810198  0.66556174  0.201000423  
## gdp_growth -0.037259418 -0.2681020  1.00000000 -0.26111105  0.015524179  
## urban_pop  0.382525283  0.6655617 -0.26111105  1.00000000  0.096537005  
## pop_dens   0.008620233  0.2010004  0.01552418  0.09653700  1.000000000  
## educ_exp   0.011627608  0.1458641 -0.10369403  0.06996132 -0.097466079  
## population 0.552621606 -0.3469278  0.08247239 -0.11629183 -0.113163291  
## pop_growth -0.182845130 -0.4287091  0.19406182 -0.22386803 -0.137773326  
##           educ_exp population pop_growth  
## gdp      0.01162761  0.55262161 -0.1828451  
## gdp_xcap  0.14586405 -0.34692778 -0.4287091  
## gdp_growth -0.10369403  0.08247239  0.1940618  
## urban_pop  0.06996132 -0.11629183 -0.2238680  
## pop_dens  -0.09746608 -0.11316329 -0.1377733  
## educ_exp   1.00000000 -0.09317277 -0.1374265  
## population -0.09317277  1.00000000  0.2029031  
## pop_growth -0.13742646  0.20290314  1.0000000
```

```
corrplot.mixed(corr.res, upper="ellipse", tl.cex = 0.6)
```



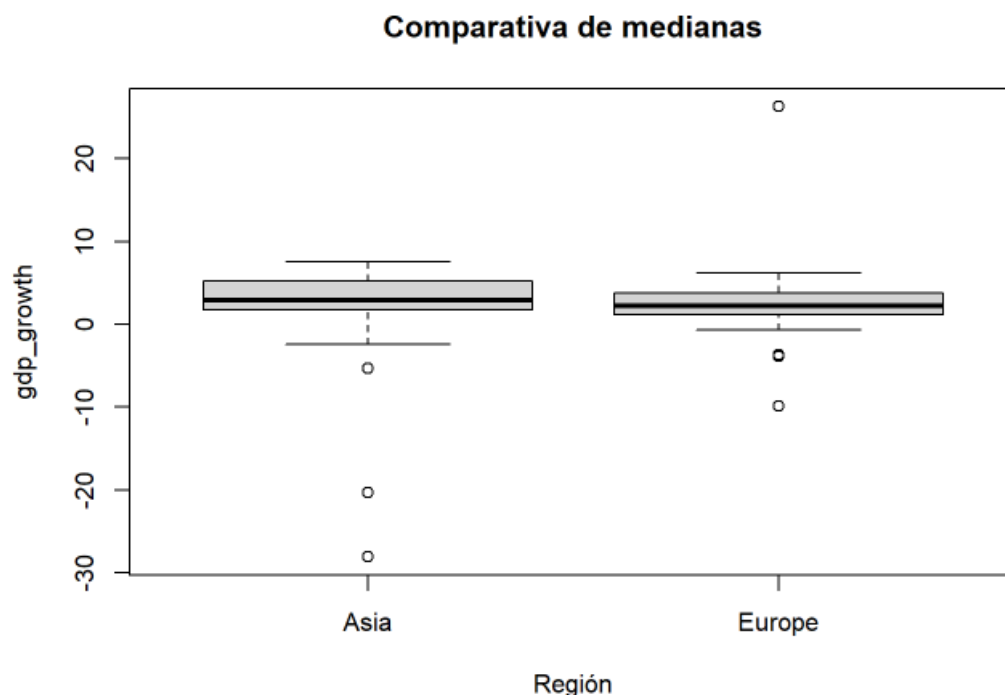
Gráficos del test de Mann-Whitney

Para comprobar si existen diferencias de medianas a un nivel visual, hemos decidido representar los Box-Plots de ambos conjuntos de datos.

Ilustración 18. Gráficos y tablas - Test de Mann-Whitney

```
# Comprobemos visualmente la diferencia de medianas:

country_profiles_europe$Region <- "Europe"
country_profiles_asia$Region <- "Asia"
country_profiles_eurasia = rbind(country_profiles_europe, country_profiles_asia)
country_profiles_eurasia$Region <- factor(country_profiles_eurasia$Region)
plot(country_profiles_eurasia$Region, country_profiles_eurasia$gdp_growth, main = "Comparativa de medianas", xlab = "Región", ylab = "gdp_growth")
```



En efecto, la mediana de crecimiento del PIB en Asia parece ser ligeramente mayor, si bien no se trata de una diferencia significativa. Es por ello, que el test de Mann-Whitney ha considerado no rechazar la hipótesis nula. Reforzamos entonces la conclusión de que ambas muestras no guardan una diferencia estadísticamente significativa.

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

A través del análisis de regresión lineal múltiple, hemos comprobado que todos los sectores económicos (agricultura, industria y servicios) contribuyen al incremento del PIB per cápita, si bien muy tímidamente como hemos podido comprobar en la representación gráfica del ejercicio 5. Por el contrario, el porcentaje de empleados de cada sector no contribuye a la explicación de esta variable.





Por otro lado, en cuanto a variables demográficas, la única que parece tener una influencia significativa en la riqueza de un país (aparte del incremento absoluto de la población), es el **porcentaje de población urbana**. Por otro lado, hemos comprobado por ejemplo cómo el **gasto en educación de un país no se traduce necesariamente en una mayor riqueza**.

Por último, **no podemos concluir que existan diferencias significativas en el crecimiento de la riqueza entre países desarrollados y países en vías de desarrollo**.

Por tanto, concluimos que, en general, **es muy difícil modelar los comportamientos que debe tener un país para aumentar su riqueza**. Si aumenta la aportación al VAB de sus sectores económicos, así como el porcentaje de población urbana, puede darse el caso de que aumente el PIB por cápita. Sin embargo, los modelos detrás de estas indagaciones no son perfectos, por lo que no deben ser tomadas como irrefutables.

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.

Código adjunto en el repositorio de GitHub, World_GDP_Patterns

Contribuciones	Firma
Investigación previa	Dionisio González (D.G.), Carlos Lavado (C.L.)  
Redacción de las respuestas	Dionisio González (D.G.), Carlos Lavado (C.L.)  
Desarrollo de código	Dionisio González (D.G.), Carlos Lavado (C.L.) 