Análisis e Interpretación de Datos

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS MASIVOS / VISUAL ANALYTICS AND BIG DATA

Miller Janny Ariza Garzón

Reflexiones Actividad 1

unir LA UNIVERSIDAD EN INTERNET

Universidad Internacional de La Rioja

Fechas de entrega próximas actividades

- Actividad Grupal: 27 de enero de 2023 (hoy dice 20 de enero)
- Laboratorio: 17 de febrero de 2023 (hoy dice 13 de febrero)

Reflexiones más importantes Actividad 1

Estadística descriptiva: información a partir del procesamiento numérico de datos.	Descripción	Puntuación máxima (puntos)	Peso %
Criterio 1	Definición del problema y referencias	2	20
Criterio 2	Elección de base de datos	2	20
Criterio 3	Análisis descriptivo numérico	2	20
Criterio 4	Análisis descriptivo gráfico	2	20
Criterio 5	Discusión y conclusiones de los resultados	2	20
		10	100 %

Definición del problema y referencias

Sea un problema, con respaldo a partir de citas, componente temporal, periodicidad, componente espacial, mencione las variables o constructos más relevantes a analizar, alcanzable y pertinente.

Impacto de XXX en YYY, en CONTEXTO (Espacio, tiempo, …)
Comparación de XXX por región en los periodos…
Análisis de la evolución de XXX en el CONTEXTO de …

Se soporta (justifica) con citas de autores que han trabajado en algo similar o que contextualizan el problema

El problema no es el objetivo.

Reflexiones más importantes Actividad 1

Elección de datos

- Mencionar las fuentes (citas)
- Se describe el conjunto: tamaño, descripción de las variables a estudiar, significados de los labels.
- Justificación de las variables elegidas en el contexto del problema.
- Si fue necesario preprocesar, se menciona brevemente pero no se detalla el paso a paso.

Estrategia de análisis

Se detalla la manera (estrategia) como se analizan los datos y como se relaciona esa estrategia con el problema estudiado, para entenderlo o para darle solución.

Ej, Comparación de medias, dispersiones, medianas y medias robustas, ya que permiten ….
Comparación de las diferentes distribuciones a partir de densidades, histogramas y boxplots, ya que permiten ….

Reflexiones más importantes Actividad 1

Elección descriptivo numérico

Es obligatorio

Análisis descriptivo gráfico

Es obligatorio

- Tablas y gráficos bien presentados.
- Citar dentro del texto.
- Títulos y estructura de gráficos y tablas según normas APA

Reflexiones más importantes Actividad 1

Discusión y conclusiones de los resultados

- Es necesario un apartado de Discusión y conclusiones
- No solo describir los resultados. Hay que asociarlos con el problema estudiado. Que se gana o se entiende con los resultados encontrados, en el contexto del problema.
- Mencionar limitaciones.
- Mencionar futuros caminos de trabajo y análisis.

Ej.
Los resultados numéricos de XX, al comparar por región, nos muestra que …. Es por esto que …. Tal como lo menciona PPP(2022)

Al contrario de lo que afirmaba PPP(2021) y ARRR(2020), en este estudio se encontró….

Reflexiones más importantes Actividad 1

Otros

- Incluir citas académicas
- Citar, no es suficiente con poner las referencias bibliográficas
- Solo se referencia lo que se cita
- La bibliografía siempre es necesaria en un documento
- Normas APA
- Dedicarle tiempo suficiente
- Revisar luego de escribir. Leer antes de entregar
- El archivo de R o Python es necesario
- El archivo latex es necesario si se ha usado
- El archivo de Markdown es necesario si se ha usado
- El documento paper a evaluar es el que se entrega en word o pdf.
- Un trabajo no es una lista de tablas y gráficos.

Reflexiones más importantes Actividad 1

Ejemplos de problemas con dificultades en su definición:

"El coronavirus (COVID-19) es una emergencia de salud pública que ha cambiado la vida de toda la población con el confinamiento obligatorio en todos los países (Berho y Beccaria, 2020). Se quiere analizar como impactó el confinamiento por el COVID-19 en los estudiantes -de varias instituciones educativas en la Región de la Capital Nacional de Delhi (NCR)- en su tiempo dedicado al aprendizaje, los medios usados para poder recibirlas y cuánto tiempo les dedicaban a las redes sociales y a la TV."

"El presente trabajo va analizar el número de muertes diarias de COVID-19 considerando el sexo, edad y ubicación geográfica del Perú, con ello determinaremos cuál de las variables a influenciado en el fallecimiento diario. El tiempo a considerar empieza del 03/03/2020 hasta 11/12/2022(Minsa-Perú,2022)."

Reflexiones más importantes Actividad 1

Ejemplos de problemas con dificultades las tablas:

	Frec_abs	Frec_rel_%
NA	125985	12.014877
8/06/2021 0:00	25446	2.426722
10/06/2021 0:00	23920	2.281191
15/06/2021 0:00	23815	2.271178
11/06/2021 0:00	21687	2.068235
...	...	...
14/03/2020 0:00	1	0.000095
1/03/2020 0:00	1	0.000095
23/03/2020 0:00	1	0.000095
24/03/2020 0:00	1	0.000095
23/12/2021 0:00	1	0.000095

524 rows × 2 columns

Tabla 2. Representación de reconocimientos de síntomas

Figura 1

Frecuencias en la variable edad

```
intervaloEdad2
   [0,6]   (6,12]  (12,18]  (18,24]  (24,30]  (30,36]  (36,42]  (42,48]  (48,54]  (54,61]  (61,67]  (67,73]
   11576    22396    39630    86290   159724   206651   191089   102079    73378    62551    37063    24180
 (73,79]  (79,85]  (85,91]  (91,97] (97,104] (104,110] (110,116]
   16204     9616     4538     1352      212       13        2
```

Nota. La Figura 1 contiene la información de frecuencias en base a la variable edad

Sentimiento tras COVID		hombre	mujer	total
Temor a enfermar	Si	682	773	1455
	No	545	461	1006
	Total	1227	1234	2461
Dolor por perdida de un familiar	Si	437	439	876
	No	789	795	1584
	Total	1226	1234	2460
Preocupación por perdida de empleo	Si	345	412	757
	No	880	822	1702
	Total	1225	1234	2459
Inquietud por limitar contactos y relaciones	Si	860	900	1760
	No	364	327	691
	Total	1224	1227	2451
Miedo por perdida de empleo	Si	557	685	1242
	No	667	546	1213
	Total	1224	1231	2455
Intranquilidad por no afrontar los gastos	Si	462	549	1011
	No	760	686	1448
	Total	1222	1235	2457
Miedo de no recuperar vida anterior	Si	629	798	1427
	no	591	490	1021
	Total	1220	1228	2448
Miedo por no poder emprender proyectos	Si	553	652	1205
	No	627	580	1252
	Total	1225	1232	2457
Inquietud y temor ante el futuro	Sii	744	908	1652
	No	472	322	794
	Total	1216	1230	2446

Tabla 1: Tablas cruzadas de las variables de sentimientos negativos en función del sexo.

Reflexiones más importantes Actividad 1

Ejemplos de problemas con dificultades las tablas:

Tomemos como referencia el país España como ejemplo del análisis, pudiéndose aplicar en el resto de países de la Base de Datos, atendiendo a los modelos estadísticos anteriormente indicados, siendo la primera la media:

```
> mean(bdmuertes$Spain)
[1] 9016.514
```

Como se puede observar, la media semanal de muertes en España entre 2020 y 2021 es de 9.016,51

Respecto a la desviación típica se puede ver que es de 2.208,16 muertes:

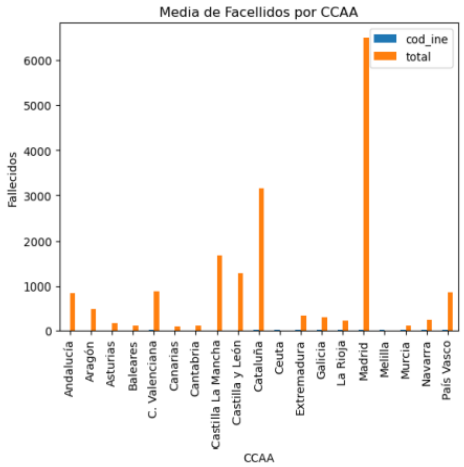
```
> sd(bdmuertes$Spain)
[1] 2208.166
```

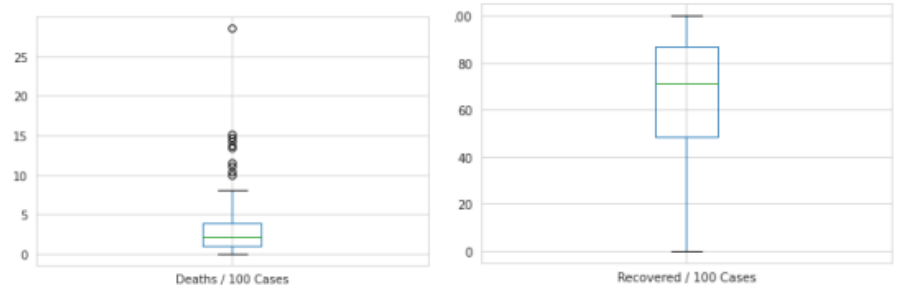
Este valor es bastante elevado dado que las muertes tienen una dispersión respecto a la media de 2.208,16.

Reflexiones más importantes Actividad 1

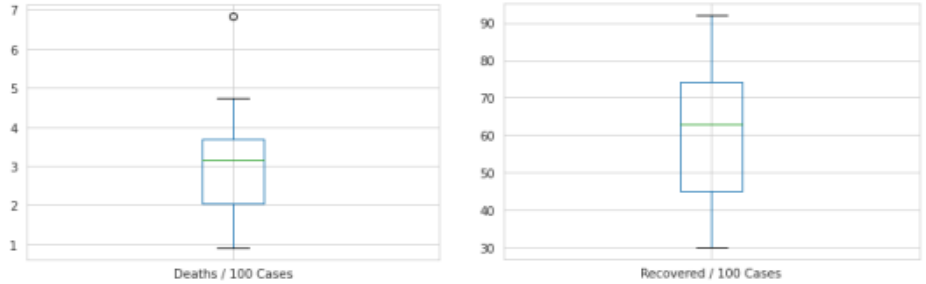
Ejemplos de problemas con dificultades en los gráficos:

Figura 1. *Media de fallecimientos por CCAA*





Gráfica 4: Porcentaje de muertes y recuperados a nivel mundial



Gráfica 5: Porcentaje de muertes y recuperados a nivel Latinoamérica

Reflexiones más importantes Actividad 1

Ejemplos de problemas con dificultades en los gráficos:

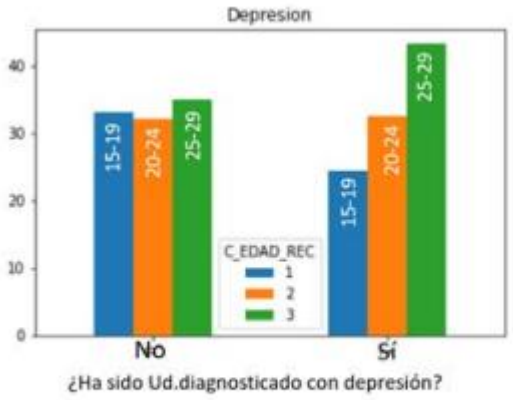


Gráfico 4. ***Comparativa por grupos de edad***

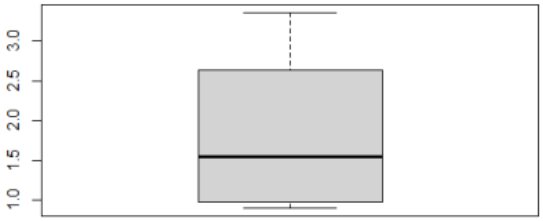
P58_1	0	1	All
C_EDAD_REC			
1	27.658143	3.970390	31.628533
2	26.716016	5.316285	32.032301
3	29.273217	7.065949	36.339166
All	83.647376	16.352624	100.000000

Tabla 4. ***Tabla de contingencia grupo edad y Pregunta*** P58_1

Reflexiones más importantes Actividad 1

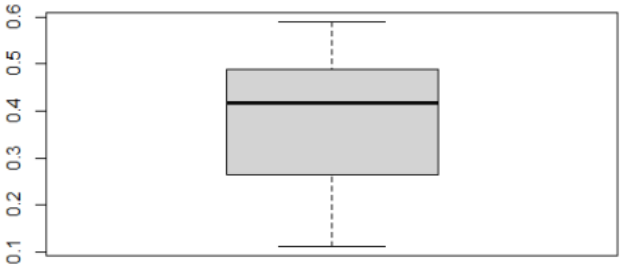
Ejemplos de problemas con dificultades en los gráficos:

Figura 1: Boxplot de la tasa de defunciones por contagiados de COVID-19 en el 2021



Fuente : Elaboración propia

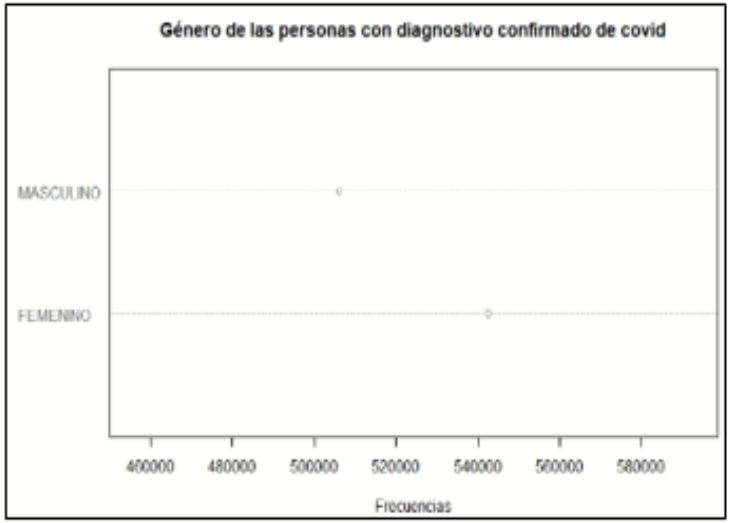
Figura 2: Boxplot de la tasa de defunciones por contagiados de COVID-19 en el 2022



Reflexiones más importantes Actividad 1

Ejemplos de problemas con dificultades en los gráficos:

Figura 3
Gráfico de frecuencias con la variable sexo



Reflexiones más importantes Actividad 1

Ejemplos de problemas con dificultades en los gráficos:

Figura 6. Diagramas de caja pretest y postest

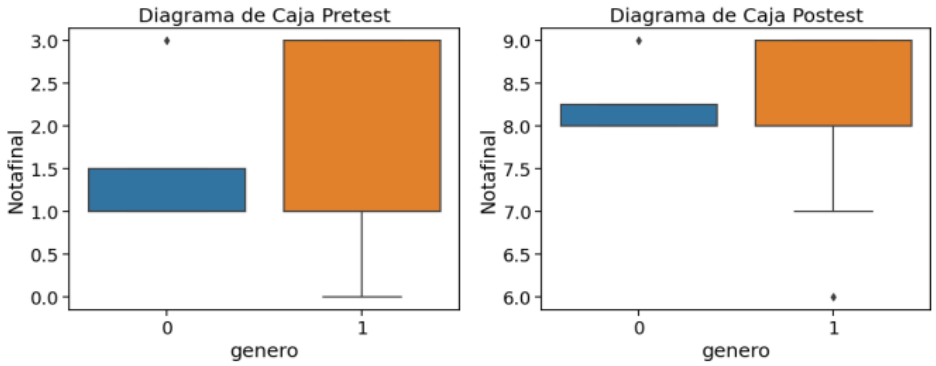


Figura 7. Contraste de hipótesis shapiro-wilk pretest

```
##Variables normales
datashapiroNorm
```

✓ 0.1s

	Variable	Valores P	Concepto
0	edad	0.451883	Es una variable Normal
6	P1	1.000000	Es una variable Normal
11	P6	1.000000	Es una variable Normal
12	P7	1.000000	Es una variable Normal
14	P9	1.000000	Es una variable Normal
15	P10	1.000000	Es una variable Normal

```
##Variables no normales
datashapiroNNorm
```

✓ 0.1s

	Variable	Valores P	Concepto
1	genero	0.000056	No es una variable Normal
2	estrato	0.008703	No es una variable Normal
3	nestudio	0.035424	No es una variable Normal
4	marcacel	0.038763	No es una variable Normal
5	gamacel	0.012704	No es una variable Normal
7	P2	0.000018	No es una variable Normal
8	P3	0.000004	No es una variable Normal
9	P4	0.000018	No es una variable Normal
10	P5	0.000018	No es una variable Normal
13	P8	0.000018	No es una variable Normal
16	Notafinal	0.004689	No es una variable Normal

Figura 9. Alpha de Cronbach instrumento de medición

```
##Analisis de confiabilidad del instrumento Alpha de Cronbach
import pingouin as pg

pg.cronbach_alpha(data=diagIni)

##Confianza del instrumento 0.65 Moderado
```

✓ 0.7s

```
(0.05832855230445609, array([-0.869,  0.657]))
```

Reflexiones más importantes Actividad 1

Ejemplos de problemas con dificultades en los gráficos:

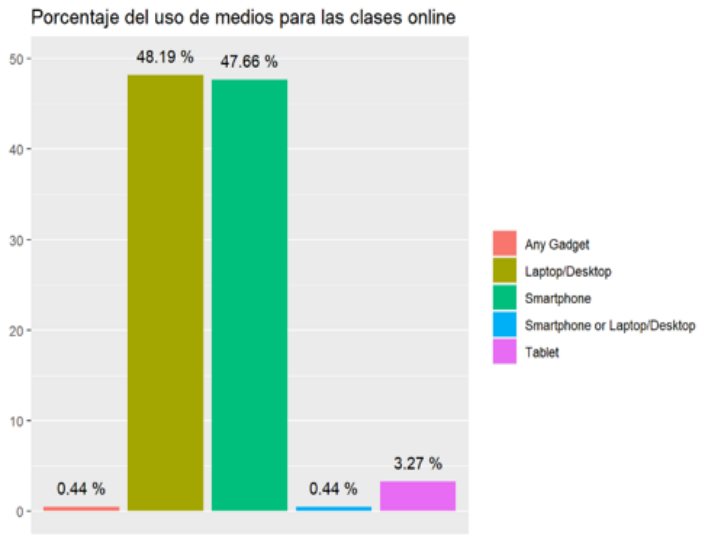
Tabla 1

Distribución de la muestra según los diferentes medios usados para el aprendizaje online

Medios usados para el aprendizaje online	Frecuencia	Porcentaje
Laptop/Desktop	545	48,19
Smartphone	539	47,66
Tablet	37	3,27
Any Gadget	5	0,44
Smartphone or Laptop/Desktop	5	0,44
Total	**1131**	

Figura 1

Porcentaje de cada medio (equipo) que se utilizó para el aprendizaje online.



Fuente. Elaboración Propia

Reflexiones más importantes Actividad 1

Una manera de presentar las variables a utilizar en el análisis:

TABLE 1. **Description of the explanatory variables according to the information in Kaggle.**

Variable	*Description*
Categorical variables	
emp_length	Employment length. Current employment time in years categorized by LC into 12 categories, including the no information category.
experience_c	Previous credit experience with LC (binary).
purpose	Purpose of the loan provided by the borrower. It has 14 possible values: car, credit_card, debt_consolidation, educational, home_improvement, house, major_purchase, medical, moving, other, renewable_energy, small_business, vacation, wedding.
home_ownership	Home ownership status provided by the borrower during the registration process. Categories defined by the entity: Mortgage, rent, own, other (other, none and any).
addr_state	State in the US provided by the borrower in the loan application.
Quantitative variables	
revenue	Yearly income self-reported in the registration process.
dti_n	Debt ratio for the group of applicants for obligations excluding mortgages. Monthly information. Income self-reported.
loan_amnt	Amount of credit requested by the borrower.
fico_n	Credit bureau score. Defined between 300 and 850, reported by Fair Isaac Corporation as a summary risk measure based on historical credit information reported at the time of application.

Reflexiones más importantes Actividad 1

Ejemplos de problemas con dificultades en las citas:

Incorrecta:

A. Ruiz-García, F. Vitelli-Storellib, A. Serrano-Cumplidoc, A. Segura-Fragosod, A. Calderón-Monteroe, R.M. Mico-Pérezf, A. Barquilla-Garcíag, Á. Morán-Bayónh, M. Linaresi, V. Olmo-Quintanaj, V. Martín-Sánchez (2022). Tasas de letalidad por SARS-CoV-2 según Comunidades Autónomas durante la segunda onda epidémica en España. Revista de Medicina de Familia ELSEVIER

Correcta:

Ruiz-García, A., Vitelli-Storelli, F., Serrano-Cumplido, A., Segura-Fragoso, A., Calderón-Montero, A., Mico-Pérez, R. M., ... & Martín-Sánchez, V. (2022). Tasas de letalidad por SARS-CoV-2 según Comunidades Autónomas durante la segunda onda epidémica en España. Medicina de Familia. SEMERGEN, 48(4), 252-262.

www.unir.net