Analítica de datos aplicada a estudios sobre desarrollo

Segundo examen

Docentes:

Giovany Babativa-Márquez

Javier García-Estévez

Instrucciones

El trabajo puede ser desarrollado como máximo en grupos de tres (3) personas. Debe entregar el script reproducible o archivo de R utilizado para generar los resultados, así como un documento de máximo *10 páginas* y debe estar conformado como mínimo por las siguientes secciones en un formato de artículo científico[1]:

- Resumen
- Introducción
- Materiales y métodos
- Resultados
- Conclusiones
- Bibliografía

Su documento debe incluir los resultados del desarrollo de los ejercicios propuestos más adelante, estos se han propuesto con el fin de que sirvan para generar una estructura retórica. Es decir, no debe entregar por separado el desarrollo de cada ejercicio, sino que todos estos elementos deben contribuir para construir el documento que debe entregar.

La fecha límite para la entrega será el ***15 de mayo del 2024*** y se debe cargar por la plataforma de Bloque Neón. *Los trabajos enviados por correo electrónico no serán considerados para su evaluación*.

[1] Cada sección y el script reproducible servirán como rubrica de evaluación, en cada sección del documento se valora, entre otros, la calidad de su contenido desde su relevancia, capacidad de análisis, uso de recursos bibliográficos y retórica; en el script se valora que funcione en todas las líneas al ser ejecutado y que reproduzca las salidas de forma correcta.

Tenga en cuenta que:

1. Se hará una comparación entre los trabajos y de encontrar una alta similaridad entonces **la nota será dividida entre el número de personas involucradas**.
2. Si el documento tiene más de las páginas permitidas (sin contar la bibliografía) **tendrá una penalidad de 5 décimas por hoja adicional**.

Entregables

1. Documento en formato Word o PDF tipo artículo académico siguiendo las instrucciones mencionadas previamente.

2. Script de R o archivo .Rmd que permita reproducir los resultados.

Desarrollo

El Proyecto de Opinión Pública Latinoamericana (LAPOP por sus siglas en inglés), es una encuesta realizada desde el año 2004 que busca medir los valores, actitudes y comportamientos democráticos. Para el año 2015 ya se realizaba en 28 países con un tamaño de muestra de más de 50.000 encuestas y considera un diseño muestral probabilístico en cada país. La ronda de 2021, corresponde al último estudio realizado y se llevó a cabo en 22 países con más de 64.000 encuestas.

Para este examen se han descargado los datos de 2993 encuestas realizadas en Colombia en la ronda 2021 y el cuestionario utilizado, los cuales fueron descargados de la página de LAPOP en este enlace. Asimismo, los estudiantes pueden acceder a los materiales de forma directa desde el repositorio de GitHub de este curso:

- *Ficha técnica*: haga clic aquí
- *Cuestionario*: haga clic aquí
- *Conjunto de datos en formato stata*: haga clic aquí

3.1 Ejercicio 1

Este punto le debería servir para dar un contexto del problema y poder escribir la introducción de su documento.

Imagine que se le ha solicitado construir tres (3) índices:

- Indice de confianza en instituciones
- Indice de democracia
- Indice de valores antidemocráticos

Para cada índice defina el concepto que desea medir argumentando su enfoque desde referencias bibliográficas y realice las citas de forma apropiada en su documento.

3.2 Ejercicio 2

Esto punto debería ayudarle a complementar su sección de materiales y métodos.

Descargue el cuestionario desde el enlace señalado previamente, e identifique las variables que usaría en cada caso para construir los tres (3) índices.

3.3 Descripción de los datos

El siguiente paso consiste en construir la base de datos. Para ello se llevaron a cabo algunas tareas que permiten el preprocesado para importar, ordenar y transformar las variables que serán relevantes en nuestro análisis, el script utilizado puede ser descargado de acá.

El conjunto de datos obtenido, luego del preprocesamiento, contiene 16 variables que fueron construidas desde las preguntas de la encuesta LAPOP, y en el que finalmente se tuvieron en cuenta 2.971 encuestas que pueden ser descargadas acá. A continuación se describen las variables utilizadas.

Nombre	Descripción	Preguntas
just_golpe	Circunstancias en que se justificaría que los militares de este país tomen el poder por un golpe de Estado.	*jc13, jc13covid*
just_cierre_cong	Justificación para que el presidente del país cierre el Congreso y gobierne sin Congreso	*jc15a*
conf_gobiero_nal	Mide en una escala del 1 al 4 la confianza en que el gobierno nacional hace lo correcto.	*anestg* en escala invertida
conf_instit	Mide en una escala del 1 al 7 el nivel de respeto a las instituciones políticas en el país del encuestado.	*b2*
conf_alcaldia	Mide en una escala de 1 a 7 el nivel de confianza en la Alcaldía.	*b32*
conf_elecciones	Mide en una escala de 1 a 7 el nivel de confianza en las elecciones	*b47a.*
conf_policia	Mide en una escala de 1 a 7 el nivel de confianza en la policía	*b18*
conf_medios	Mide en una escala del 1 al 7 el nivel de confianza en los medios de comunicación	*b37*
conf_fuerzas_mil	Mide en una escala del 1 al 7 el nivel de confianza en las fuerzas militares.	*b12*
sat_democracia	Variable dicotómica que mide el nivel de satisfacción con la democracia de los encuestados.	*pn4* en escala invertida

prot_derechos	Mide en una escala del 1 al 7 la protección de los derechos básicos del ciudadano por parte del sistema político colombiano	b3
orgullo_sistema	Mide en una escala del 1 al 7 qué tanto se siente usted orgulloso de vivir bajo el sistema político colombiano	d4

Para realizar los análisis cargue los paquetes necesarios y el conjunto de datos, recuerde que puede hacerlo así:

```
rm(list = ls())

options(scipen = 999)
library(pacman)

p_load(tidyverse, janitor, corrplot, haven,
       devtools, FactoMineR, factoextra,
       ggcorrplot, GGally)

url <- "https://github.com/jgbabativam/AnaDatos/raw/main/datos/df_colombia.dta"

datos <- read_dta(url)
```

Considerando la naturaleza de las variables que fueron elegidas, describa por qué un análisis de componentes principales es apropiado y describa brevemente en qué consiste el método (esto le debe permitir complementar su sección de materiales y métodos).

3.4 Estrategia de analítica

Los ejercicios propuestos deben permitirle generar el desarrollo de la sección de *resultados*. Siga las reglas para llevar la numeración de las gráficas y tablas, y recuerde que todas las que decida usar deben llevar una descripción o interpretación dentro de su documento. Use una retórica que facilite la lectura de los resultados, en donde el lector pueda llevar un hilo conductor que sea atractivo y amable - trate de ir contando una historia.

A partir de la base de datos construida desarrolle los siguientes ejercicios:

3. Analice la relación lineal entre las variables, para ello calcule la matriz de correlación entre las variables cuantitativas y represente el resultado con un diagrama usando el paquete `corrplot` o `ggcorrplot`. Concluya sobre estos resultados.

Realice un Análisis de Componentes Principales sobre las variables cuantitativas:

4. Discuta sobre la cantidad de información del conjunto de los datos es explicado por las dos primeras dimensiones. Presente un gráfico que lo ilustre. Ayuda use la función `fviz_screeplot()` del paquete `factoextra`.

5. Haga una gráfica del plano factorial generado por las dimensiones 1 y 2 para las variables. Concluya sobre la asociación de las variables indicando cuáles presentan fuertes correlaciones y apuntan en la misma dirección representando el mismo concepto, indique también si considera que hay variables que no se representen bien en el plano de las primeras dos dimensiones. Ayuda use la función `fviz_pca_var()` del paquete `factoextra`.

6. Use el siguiente comando para visualizar de forma simultánea a los encuestados y las variables

```
fviz_pca_biplot(res, repel = F, col.var = "black", col.ind = "gray")
```

Tenga en cuenta que `res` se refiere al objeto que contiene los resultados del PCA. De forma visual, considerando la densidad de punto grises (encuestados), ¿podría conjeturar que hay una buena confianza en las instituciones? explique.

7. A partir de los resultados, explique que representa un puntaje alto en la primera componente principal. Asigne un nombre apropiado al índice que se obtendría desde esta dimensión. Apoye su conclusión en el gráfico que resulta del siguiente comando (tenga en cuenta que `res` se refiere al objeto que contiene los resultados del PCA):

```
fviz_contrib(res, choice = "var", axes = 1, top = 10)
```

8. A partir de los resultados, explique que representa un puntaje alto en la segunda componente principal. Asigne un nombre al índice que se obtendría desde esta dimensión. Apoye su conclusión en el gráfico que resulta del siguiente comando (tenga en cuenta que `res` se refiere al objeto que contiene los resultados del PCA):

```
fviz_contrib(res, choice = "var", axes = 2, top = 10)
```

9. A partir de los resultados, explique que representa un puntaje alto en la tercera componente principal. Asigne un nombre al índice que se obtendría desde esta dimensión. Apoye su conclusión en el gráfico que resulta del siguiente comando (tenga en cuenta que `res` se refiere al objeto que contiene los resultados del PCA):

```
fviz_contrib(res, choice = "var", axes = 3, top = 10)
```

10. Construya los índices formados por las tres (3) primeras componentes principales, debe hacer uno por cada componente principal y agreguelos al conjunto de datos. Ayuda: use el siguiente código para hacerlo:

```
##### Indices

index1 <- as.data.frame(res$ind$coord[,1]) |>
          rename(score = `res$ind$coord[, 1]`) |>
          mutate(Indice1 = round(GGally::rescale01(score)*100, 1)) |>
          select(-score)
```

```
index2 <- as.data.frame(res$ind$coord[,2]) |>
          rename(score = `res$ind$coord[, 2]`) |>
          mutate(Indice2 = round(GGally::rescale01(score)*100, 1)) |>
          select(-score)

index3 <- as.data.frame(res$ind$coord[,3]) |>
          rename(score = `res$ind$coord[, 3]`) |>
          mutate(Indice3 = round(GGally::rescale01(score)*100, 1)) |>
          select(-score)

df_index <- bind_cols(datos, index1, index2, index3)
```

11. Realice los análisis descriptivos del valor de los índices por región, sexo, edad y nivel educativo. Para ello puede calcular el promedio o usar gráficas apropiadas que le permitan concluir si existe una mayor confianza en determinadas regiones, rangos de edad o niveles educativos.

12. Genere las conclusiones generales del ejercicio. Esto le debe permitir escribir la sección de conclusiones.