Prueba para los postulantes de prácticas pre-profesionales

Evaluación de conocimientos de programación y estadística

AUTHOR
Centro de Investigación Estadística ERGOSTATS

El objetivo general de los ejercicios presentados a continuación es evaluar los conocimientos en el manejo y análisis de datos con el lenguaje R (o python) de los candidatos a nuevos miembros del equipo de trabajo. Esta evaluación está destinada a los estudiantes que buscan realizar sus prácticas pre profesionales y ampliar sus conocimientos de estadística, matemática, economía y programación en lenguajes de código abierto (principalmente R). Con estos ejercicios se busca medir su comprensión del contenido de una base de datos, su habilidad para manipularla, generar estadísticas de resumen, comparar grupos dentro de los datos y analizar patrones. Para esto se utilizará la base de datos de la Encuesta Empresarial Estructural del INEC (Ecuador).

Cada ejercicio se acompañará con una serie de instrucciones y un **hint** para ayudar a los estudiantes a resolverlo. Al finalizar la evaluación, se espera tener una idea clara del nivel de conocimientos en el manejo y análisis de datos con R de los estudiantes y poder brindarles retroalimentación para que puedan mejorar en las áreas que lo necesiten.

Detalles sobre la presentación

- Para la resolución de esta prueba los estudiantes **tendrán un plazo de 36 horas** desde la recepción de este documento.
- Para la presentación de los resultados el estudiante deberá generar un `script de R` el cual deberá estar ordenado y correctamente comentado.

  > Si el estudiante conoce del uso de archivos Rmarkdown se recomienda ampliamente el uso de estos archivos.

- El nombre del `script de R` a ser enviado debe seguir las siguientes instrucciones:
  - Empezar el nombre del archivo con `solucion_`
  - Continuar con las iniciales de los dos nombres y los dos apellidos del estudiante
  - Agregar la fecha de entrega con el siguiente formato `20230420` (Para el 20 de abril de 2023)
  - Omitir espacios
  - **Ejemplo:** `solucion_avbr_20230421.R`
- Para los ejercicios que involucren generar visualización de datos:
  - Exportar las imágenes a formato `.png`
  - Revisar que las dimensiones de la imagen resultante sean legibles
  - Incluir etiquetas que faciliten la lectura del gráfico
- En caso de que el estudiante determine que no le es posible concluir alguno de los ejercicios por favor añadir un comentario que indique cual hubiera sido la estrategia para resolver el ejercicio.
- En caso de que el estudiante emplee alguna herramienta de apoyo para resolver el ejercicio señalar cual fue esta herramienta.

- Si necesitan que las preguntas sean aclaradas en algún punto por favor enviar un correo a `alexvbr@ergostats.org`
- Los archivos (script o archivo Rmarkdown e imágenes) y todo el material que el estudiante crea conveniente para su presentación deberán ser a los siguientes correos:
    - Alex Bajaña: alexvbr@ergostats.org
    - Estafanía Tapia: estafaniantm@ergostats.org

Ejercicios sobre manipulación de datos

Ejercicio 1: Comprender y manipular una base de datos

- Descargar y cargar los datos del **Tomo I** la Encuesta Empresarial Estructural del INEC (Ecuador) para el año 2021 en R.

  https://www.ecuadorencifras.gob.ec/encuesta-a-empresas/
- Verificar que la estructura de la base de datos se ha cargado correctamente.
- Seleccionar y mostrar las primeras 10 filas de la base de datos.
- Seleccionar y mostrar únicamente las variables de la base de datos que contienen información sobre el identificador del establecimiento, su actividad económica y la provincia donde se encuentra ubicado.
- **Hint:** Para cargar la base de datos en R, puedes usar la función `read_csv()` de la librería `readr` dentro de la librería `tidyverse`. Para seleccionar variables específicas, puedes usar la función `select()` del librería `dplyr`.

Ejercicio 2: Generación de estadísticas de resumen

- Calcular la media, la desviación estándar y la mediana del número de trabajadores (Personal ocupado) en los empresas de la base de datos.
- Calcular el número total de empresas que existen en cada provincia.
- **Hint:** Puedes utilizar la función `summarize()` de la librería `dplyr` para calcular estadísticas de resumen. Para contar el número de empresas en cada provincia, puedes usar la función `count()`.

Ejercicio 3: Comparación de grupos dentro de los datos

- Calcular el número de empresas en la provincia de Pichincha que pertenecen a cada una de las actividades económicas (Sección CIIU) de la base de datos.
- Calcular el número de trabajadores (personal ocupado) promedio en los empresas de la provincia de Pichincha para cada una de las actividades económicas (Sección CIIU).
- **Hint:** Puedes utilizar la función `filter()` para seleccionar únicamente los empresas de la provincia de Pichincha. Luego, puedes usar la función `group_by()` para agrupar los datos por actividad económica y la función `summarize()` para calcular estadísticas de resumen por grupo.

Ejercicio 4: Análisis de patrones

- Crear una tabla de frecuencias que muestre el número de empresas en cada provincia que tienen un número de trabajadores (personal ocupado) mayor o igual a 50.
- Calcular el porcentaje de empresas en cada provincia que tienen un número de trabajadores mayor o igual a 50.

- **Hint:** Puedes utilizar la función `mutate()` para crear una variable que indique si el número de trabajadores es mayor o igual a 50. Luego, puedes usar la función `group_by()` y `summarize()` para calcular el número y el porcentaje de empresas en cada provincia que cumplen con esa condición.

Ejercicio 5: Manipulación avanzada de una base de datos

- Crear una nueva variable en la base de datos que indique si el establecimiento tiene más de 10 años de antigüedad (En la base de datos existe la variable *Año desde que la empresa dispone de RUC*).
- Crear una nueva base de datos que contenga únicamente los establecimientos que tienen más de 10 años de antigüedad y que tienen más de 20 trabajadores.
- **Hint:** Utiliza la función `mutate()` para crear la nueva variable "antiguedad" que tome el valor de 1 si el establecimiento tiene más de 10 años de antigüedad y 0 en caso contrario. Luego, utiliza la función `filter()` para seleccionar únicamente los establecimientos que cumplen con ambas condiciones (antigüedad > 0 y número de trabajadores > 20). Finalmente, utiliza la función `select()` para seleccionar únicamente las variables indicadas.

Ejercicios sobre visualización de datos

Ejercicio 1:

- Usando la base de datos de la Encuesta Empresarial Estructural del INEC, grafica un diagrama de barras que muestre el número de empresas por provincia. Asegúrese de que las provincias aparezcan en orden alfabético en el eje x y de que el gráfico tenga un título y etiquetas de los ejes claras.
- Exportar el resultado como se señalo en las indicaciones. En caso de usar archivos Rmarkdown verificar que la impresión en el documento sea adecuada.
- **Hint:** Para crear un diagrama de barras en `ggplot2`, puede utilizar la función `geom_bar()`. Ten presente el orden de las barras para la presentación.

Ejercicio 2:

- Crea un diagrama de dispersión que muestre la relación entre el número de empresas y el valor agregado bruto por provincia. Utiliza diferentes colores para cada provinciab y asegúrete de que el gráfico tenga un título y etiquetas de los ejes claras.
- **Hint:** Para crear un diagrama de dispersión en `ggplot2`, puede utilizar la función `geom_point()`. Para asignar diferentes colores a cada provincia, utilice el argumento `color = provincia` dentro de la `función aes()`.
- Exportar el resultado como se señalo en las indicaciones. En caso de usar archivos Rmarkdown verificar que la impresión en el documento sea adecuada.

Éxitos a todos los postulantes.