

**FUNDACIÓN UNIVERSITARIA KONRAD LORENZ**

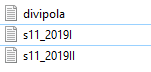
ESPECIALIZACIÓN EN ANALÍTICA ESTRATÉGICA DE DATOS

TALLER 1 BIG DATA

JOSÉ FERNANDO ZEA CASTRO

28 febrero de 2020

Para este taller se las pruebas saber del 2019 y la divipola, haga uso de dplyr y de data.table y compare los tiempos de desempeño.



1. Seleccione los estudiantes de Bogotá que presentaron las pruebas saber del primer semestre del 2019, conserve sólo las columnas del identificador del estudiante y las asignaturas evaluadas por el estudiante (no conserve el puntaje global). Crea una tabla de formato largo análoga a la que se presenta a continuación.

|  |  |  |
| --- | --- | --- |
| **ID** | **Puntaje** | **Asignatura** |
| 1 |  | Ciencias Naturales |
| . |  | . |
| . |  | . |
| . |  | . |
| 1 |  | Sociales |
| 2 |  | Ciencias Naturales |
| . |  | . |
| . |  | . |
| . |  | . |
| 2 |  | Sociales |
| . |  | . |
| . |  | . |
| . |  | . |
| n |  | Ciencias Naturales |
| . |  | . |
| . |  | . |
| . |  | . |
| n |  | Sociales |

Realice un diagrama de caja en donde compare el desempeño de las asignaturas (en R puede usar un código similar al siguiente código:

boxplot(base\_larga$Puntaje ~ base\_larga$Asignatura )

Puede correr el anterior código también con Python.

1. Use las pruebas saber consolidadas del año 2019 pero limite las columnas necesarias para este punto. Calcule el promedio del puntaje total para cada uno de los colegios de Antioquia (use el identificador del colegio para hacer la agrupación). Presente una tabla con los colegios ordenados del mejor puntaje total al menor. (Imprima los primeros registros). Realizar esta solución con sql con pandasql.
2. Use las pruebas saber consolidadas del año 2019 pero limite las columnas necesarias para este punto. Exporte los resultados obtenidos en el punto anterior a un archivo separado por comas, a un archivo de texto (txt) separado por y un archivo de Excel.
3. Punto un poco largo pero no complicado: Use las pruebas saber consolidadas del año 2019 pero limite las columnas necesarias para este punto. Integre los nombres de los departamentos de la tabla divipola a las pruebas consolidadas del año 2019. Posteriormente recodifique los departamentos a cada una de las 5 regiones (Atlántica, Andina, ver: <https://es.wikipedia.org/wiki/Categor%C3%ADa:Regiones_naturales_de_Colombia> ). Calcule el desempeño promedio del puntaje global de los estudiantes para cada una de las cinco regiones. Sugerencia: Primero realice un archivo de Excel con los códigos de los departamentos, los nombres e crea una columna con la región. Lee el archivo y realiza el pegado apropiado.
4. Use las pruebas saber consolidadas del año 2019 pero limite las columnas necesarias para este punto. Use las pruebas saber consolidadas del año 2019. Lleve a cabo el cálculo del promedio en matemáticas para las ciudades de Cali, Medellín y Barranquilla de los colegios públicos.
5. Use las pruebas saber consolidadas del año 2019 pero limite las columnas necesarias para este punto. Lleve a cabo el cálculo del promedio de los estudiantes en matemáticas para las ciudades de Cali, Medellín y Barranquilla de los colegios públicos y de más de 30 estudiantes.
6. Use las pruebas saber consolidadas del año 2019 pero limite las columnas necesarias para este punto. Seleccione los 5 mejores estudiantes de matemáticas de Cali, Medellín y Barranquilla. Ayuda: <https://stackoverflow.com/questions/29353096/add-a-sequence-number-to-each-element-in-a-group-using-python>
7. Use las pruebas saber consolidadas del año 2019 pero limite las columnas necesarias para este punto. Construya el puntaje global calculando la mediana de las asignaturas abordadas por el estudiante.

Posteriormente calcule un ranking de los 100 mejores colegios del país según este puntaje.

Ayuda: <https://www.geeksforgeeks.org/python-pandas-dataframe-median/>

1. Use las pruebas saber consolidadas del año 2019 pero limite las columnas necesarias para este punto. Calcule las correlaciones entre el puntaje de matemáticas y ciencias para cada munipio.

Ayuda: <https://stackoverflow.com/questions/45064916/how-to-find-the-correlation-between-a-group-of-values-in-a-pandas-dataframe-colu>

1. Use las pruebas saber consolidadas del año 2019. Identifique para brindar una beca a los estudiantes que se encuentran en el decil 10 de matemáticas y en el decil 10 de inglés en Bogotá y Medellín. Realice la consulta apropiada.