Universidad del Rosario

MACC
Matemáticas Aplicadas y Ciencias de la Computación

Tercera entrega proyecto Manejo de Bases de Datos 2021-I

Germán David Plazas Cayachoa, David Alfonso Oviedo Salamanca, Santiago Rodríguez Morales, Jose Gabriel Álvarez Medina.

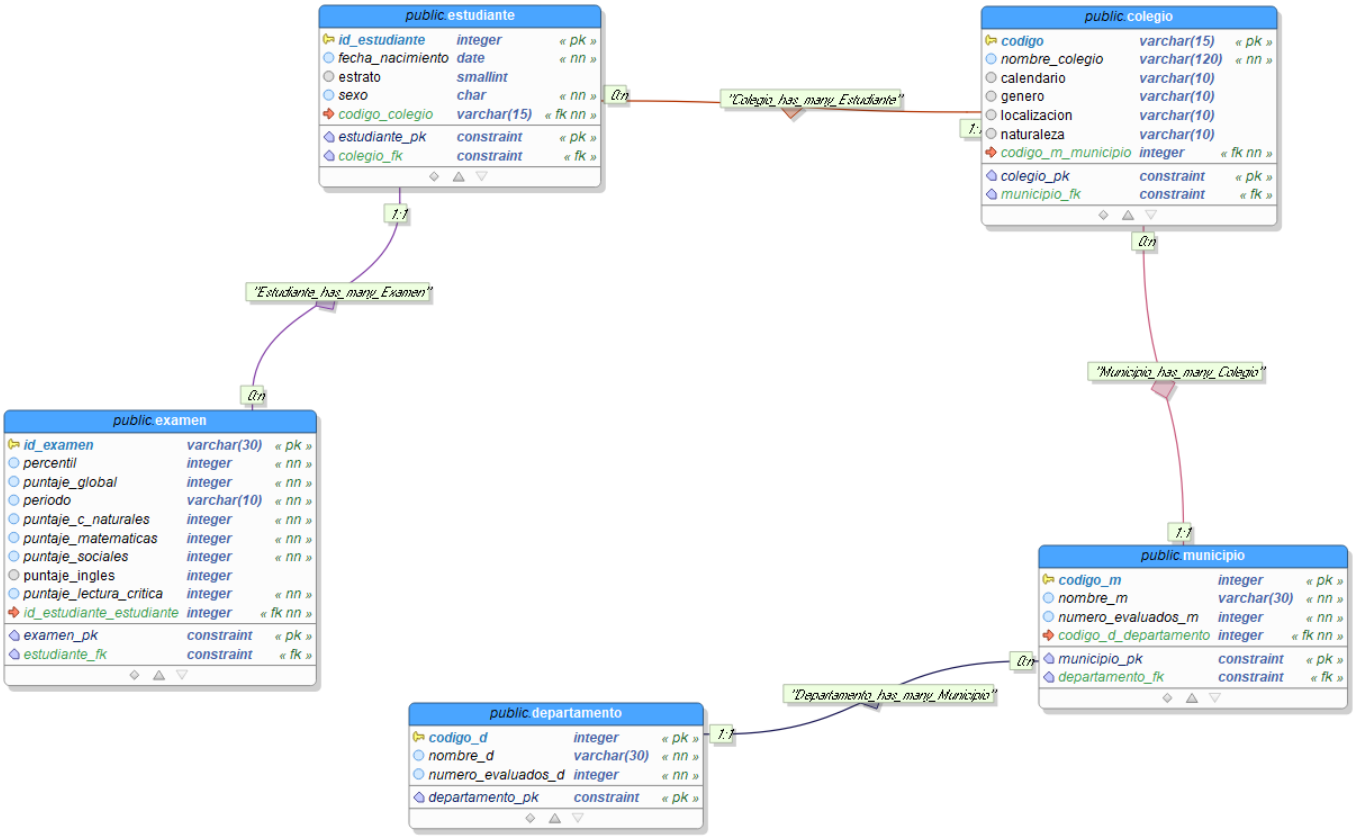
Escuela de Ingeniería, Ciencia y Tecnología, Universidad del Rosario

1. Descripción del problema

El examen de estado ICFES Saber-11 evalúa el grado de conocimiento de las áreas académicas de los estudiantes que están por finalizar la educación media. Más allá de los resultados individuales, este examen también recolecta datos de carácter socioeconómico que se pueden asociar a los resultados de algunos grupos poblacionales y reflejar ciertas fortalezas y carencias en la educación del país. Debido a lo anterior, este proyecto busca analizar y representar relaciones existentes entre algunas variables y datos obtenidos del examen.

Los datos que se usarán en este proyecto fueron obtenidos de: `https://www.datos.gov.co/Educaci-n/Saber-11-2019-2/ynam-yc42/data`, proyecto gubernamental.

2. Versión definitiva del Modelo Relacional en Tercera Forma Normal



3. Descripción de análisis desarrollados.

- Como primer análisis, se puede determinar los colegios que mas sobresalieron en las pruebas ICFES Saber 11 2019-II a nivel departamental. Aquí, se pueden realizar variedad de comparaciones que permiten concluir cómo el puntaje global promedio de los mejores colegios en departamentos más abandonados por el Estado, es muy inferior con respecto a departamentos que presentan un alta intervención socioeconómica por parte del mismo.

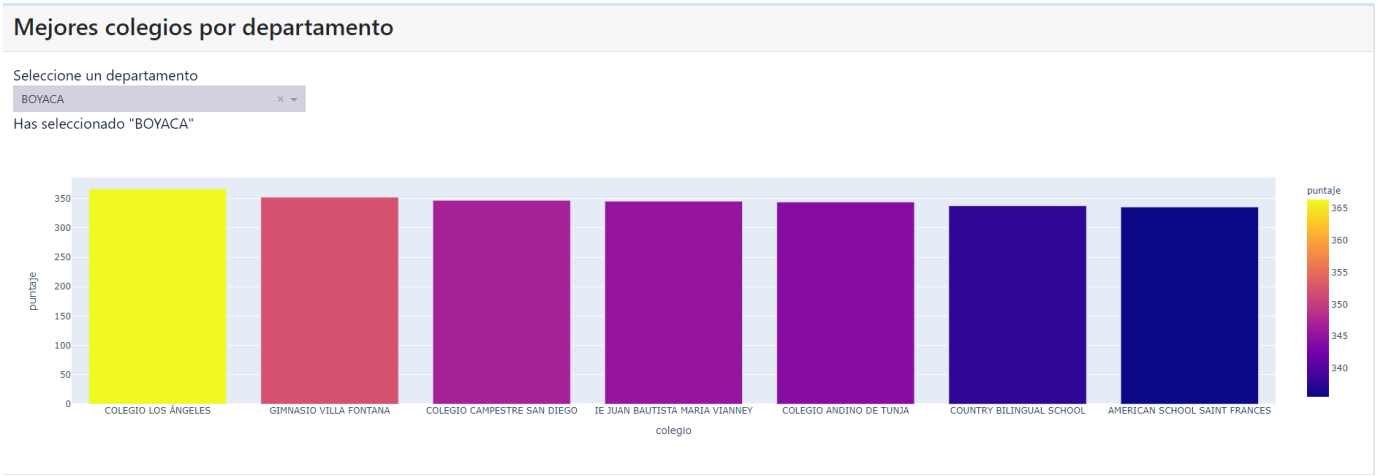


Figura 1: Diagrama de barras mejores colegios por departamento.

- Con repecto a la Figura 2, el usuario puede seleccionar el departamento al cual desea realizar el análisis, y a continuación se despliega un gráfico de barras donde se puede observar los primeros siete colegios que se destacaron en dicho departamento. En la Figura 1 se puede observar los respectivos colegios para el departamento de Boyacá. En la Figura 2, se puede observar el menú desplegable en el cual el usuario puede seleccionar uno de los 32 departamentos.

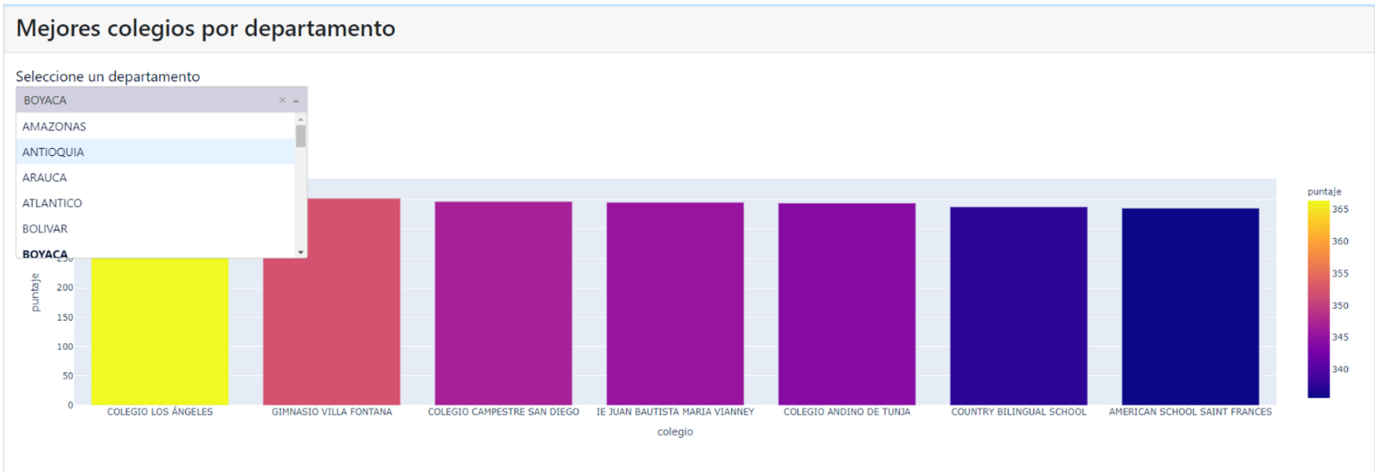


Figura 2: Selección de Departamento para el Diagrama de Barras por Departamento.

- En el siguiente análisis se deseaba realizar una comparación de los puntajes de cada materia por estrato. En la Figura 3 se desarrolla un diagrama de barras donde se evidencia cómo el promedio general de las materias va aumentando del estrato 1 al 3 en los cuales se rondan los 48 puntos por materia, por otro lado, en los estratos del 4 al 6 se rondan los 53 puntos por materia. La asignatura de sociales fue la componente que a lo largo de todos los estratos se mantuvo alta, además, en los estratos altos la componente de lectura también sobresale.

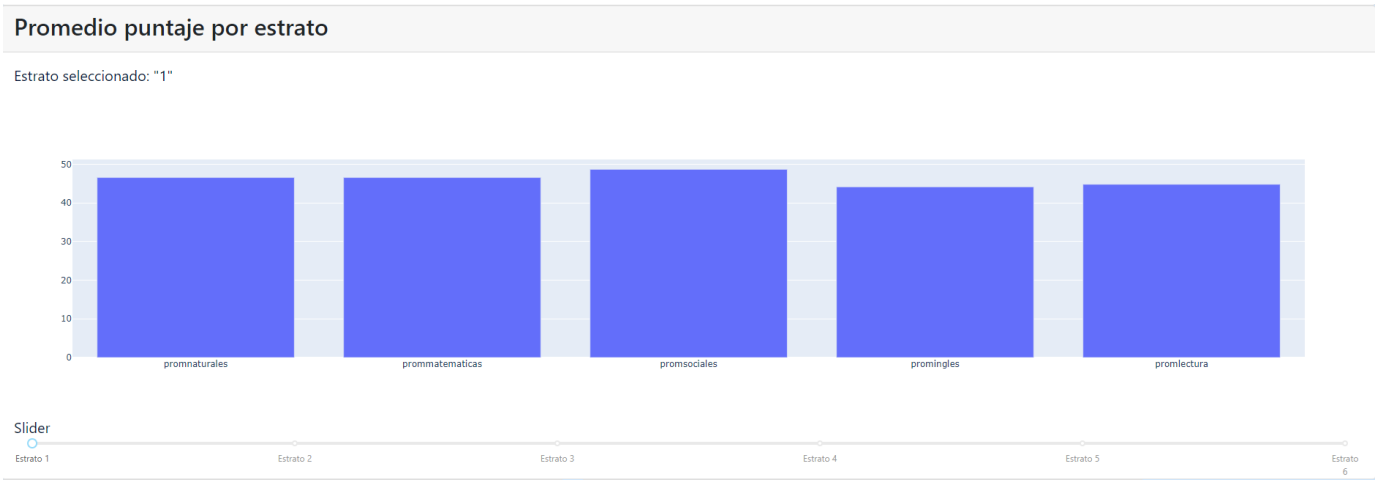


Figura 3: Slider de los puntajes de cada asignatura por estrato.

- En otro análisis, se desea conocer que tan bueno es el desempeño en el área de inglés y por ende el manejo del segundo idioma a nivel departamental de los

estudiantes. Para esto realizamos un mapa de calor dividido por departamentos en el cual se ve reflejado el puntaje promedio del exámen de inglés en las pruebas Saber 11.

Obtenemos un resultado muy interesante para nuestro análisis, en el cual concluímos que tal como se evidencia, el nivel de inglés es mayor en las áreas más pobladas del país, sobre todo en la región del centro y la ciudad capital. Esto contrasta con los bajos rendimientos de zonas alejadas o más desfavorecidas.

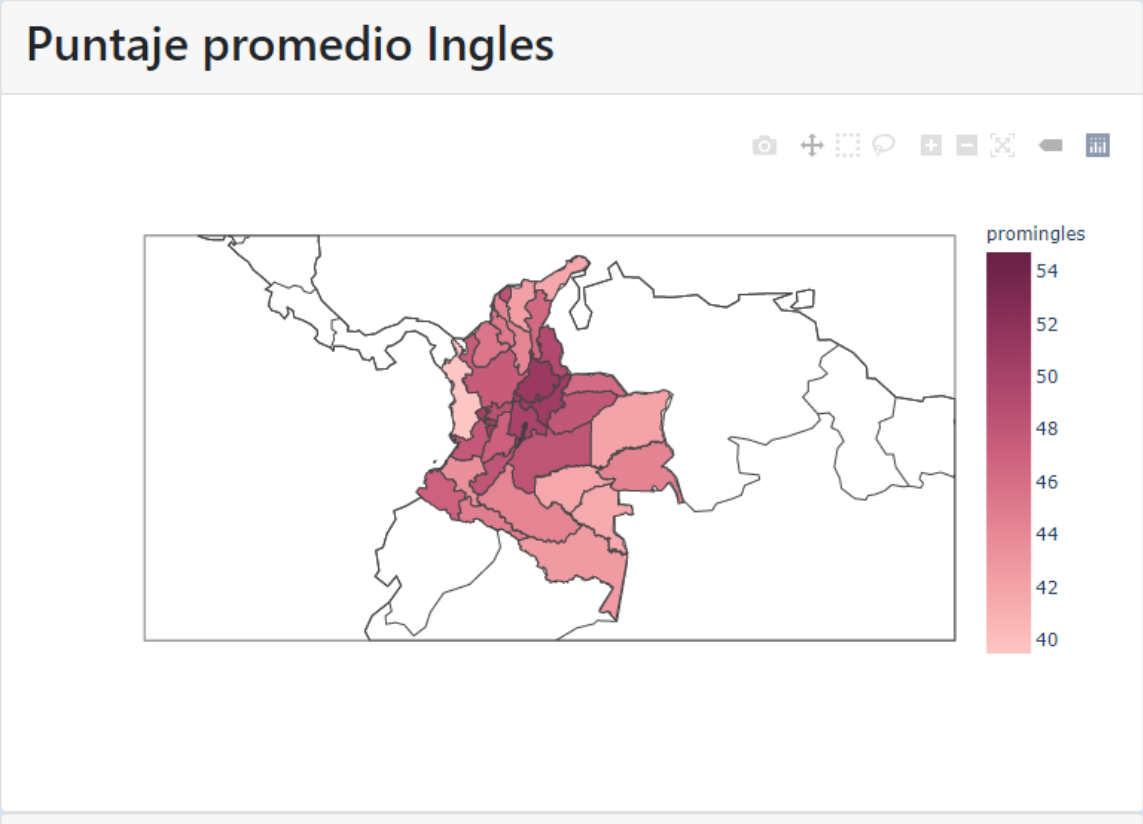


Figura 4: Mapa de calor con el puntaje promedio en las pruebas de inglés período 2019-II por departamentos

- La base de datos que se emplea en nuestro proyecto proporciona datos personales de los estudiantes que pueden reflejar características socioeconómicas del país. Debido a lo anterior, el grupo decidió buscar una representación de la distribución de los estratos de los estudiantes, teniendo como hipótesis que se encontraría una gran desigualdad. En la Figura 5 se puede visualizar un diagrama de torta que muestra la distribución de los estratos de los estudiantes. Como se puede observar, los estratos más comunes son los bajos (1,2,3),

mientras que los estratos altos (5,6) apenas superan el 2 % de la población, lo cual refleja una gran desigualdad social en la educación colombiana.

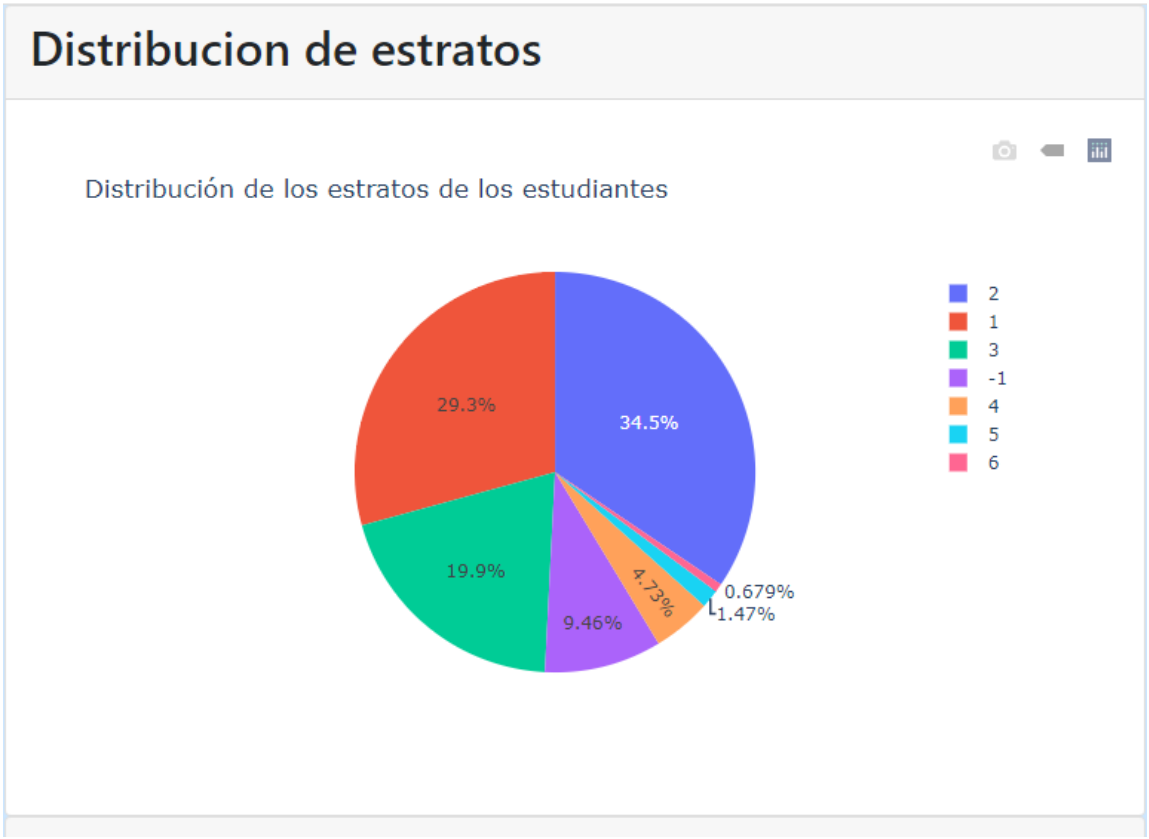


Figura 5: Diagrama de torta de la distribución de los estratos de los estudiantes.

- El grupo decidió realizar una comparación de los promedios de puntaje global en cada estrato del país. En la Figura 6 se puede visualizar un diagrama de barras en el cual destacan los promedios que sobrepasan los 250 puntos correspondiendo a los estratos 3,4 y 5, siendo el estrato 4 el de mejor desempeño. Los promedios inferiores se encuentran en ambos extremos correspondiendo a los estratos 1 y para sorpresa del grupo, el estrato 6. Con esto se concluye que los estudiantes estrato 4 pertenecen a instituciones educativas donde se preparan de mejor manera para las pruebas, de manera similar, ocurre con los estratos 3 y 5.

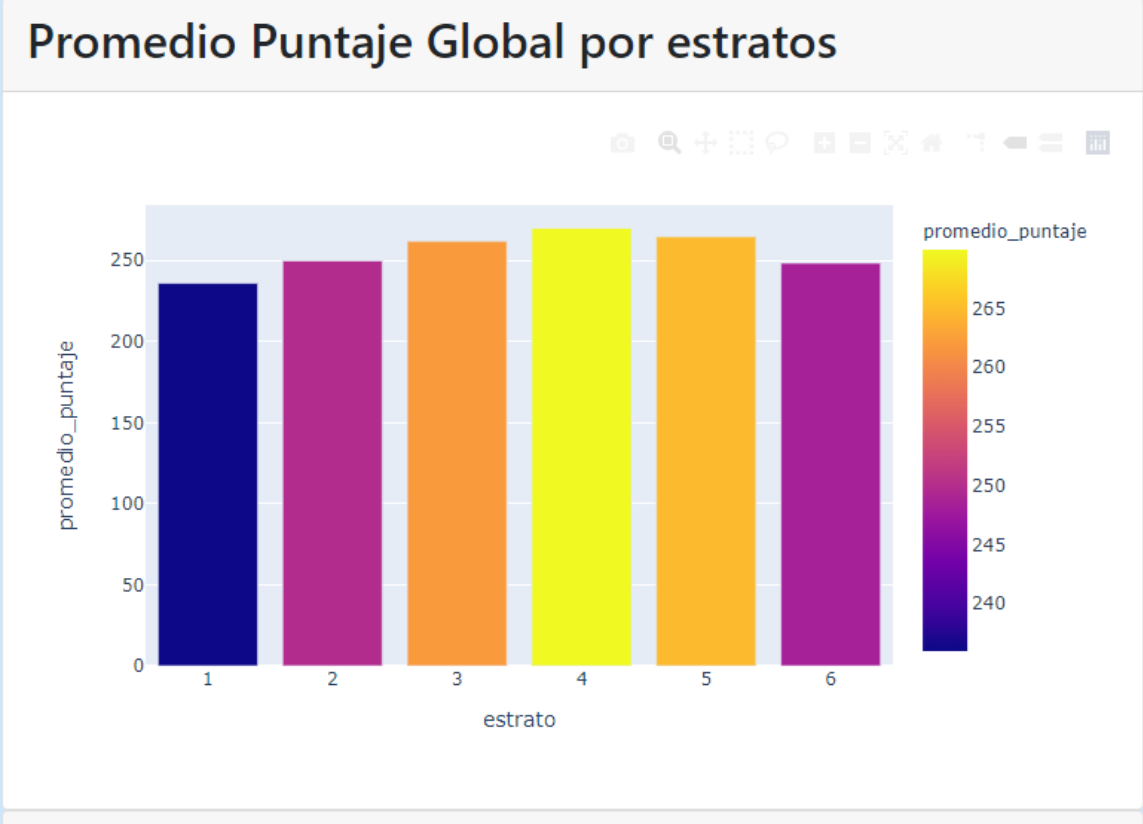


Figura 6: Diagrama de barras del puntaje global por estrato.

- El examen ICFES Saber-11 además de evaluar aspectos de educación en todo el país, también es una oportunidad para que algunas instituciones educativas adquieran prestigio en base a sus resultados, atrayendo así a que padres de familia matriculen a sus hijos. Como vemos en la Figura 7, se realiza un diagrama de barras con los promedios de los puntajes globales. De la figura podemos observar que los tres mejores colegios fueron el Centro Educativo Antares, el Liceo Campo David y la Institución Alberto Merani.

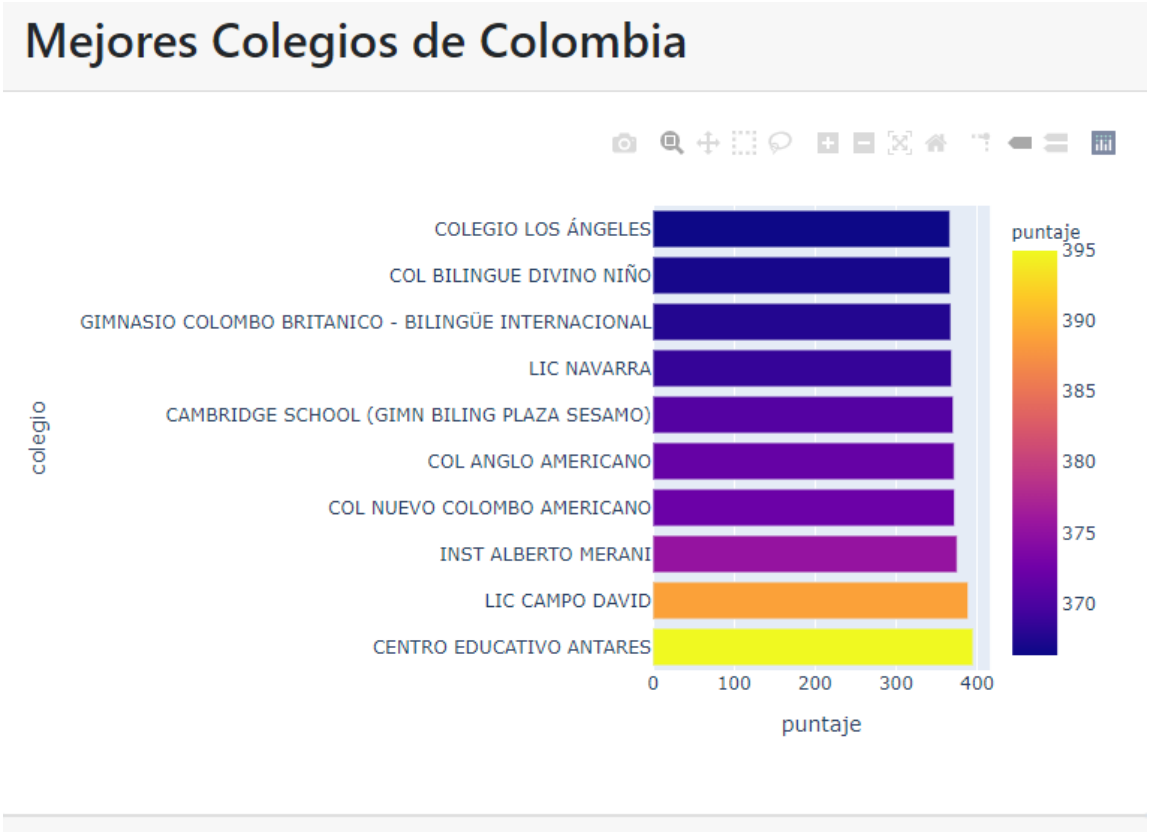


Figura 7: Diagrama de barras horizontal de los puntajes de los diez mejores colegios de Colombia

- En la Figura 8 se compara el desempeño en esta prueba entre colegios públicos y privados y se concluye que los colegios de naturaleza oficial tienen un desempeño ligeramente menor que los colegios privados, pues el promedio de puntaje global para los privados es de 263.39 mientras que el puntaje para los colegios publicos se encuentra oscilando entre los 240 y 241 puntos.

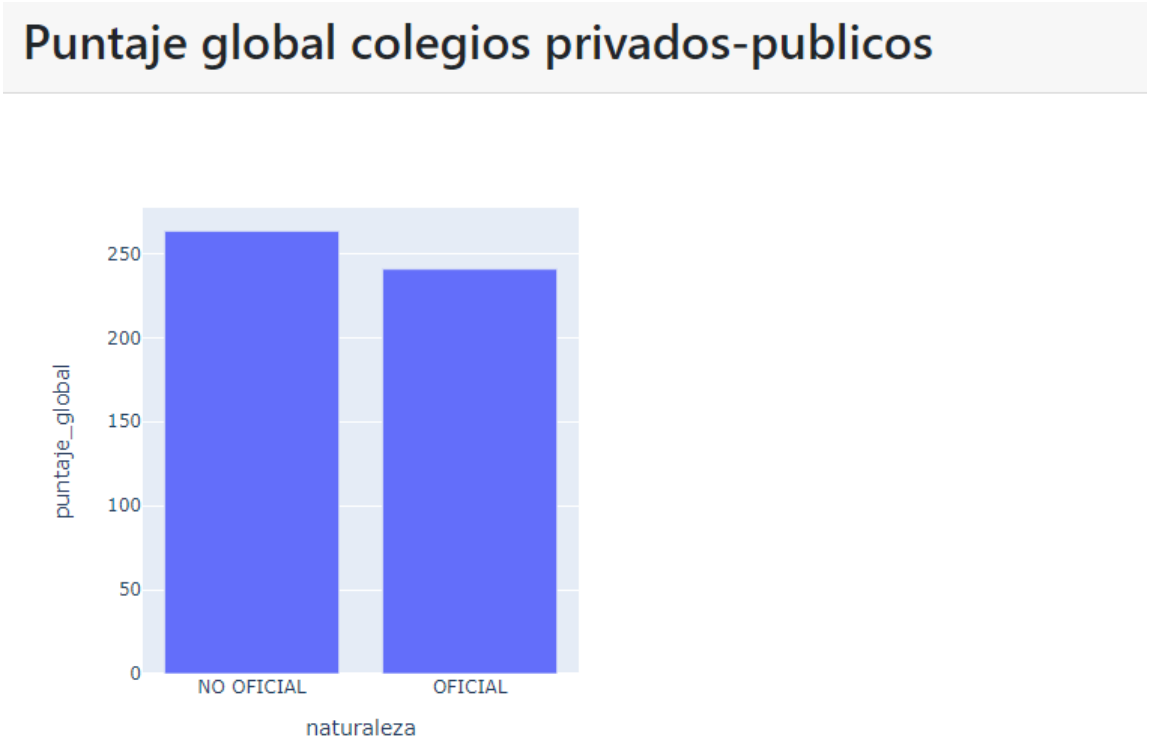


Figura 8: Diagrama de barras con el puntaje global obtenido por colegios de naturaleza oficial y no oficial

- En un análisis final, se desea evidenciar el desempeño general en las pruebas Saber 11 teniendo en cuenta el sexo del estudiante. Como se puede observar en la Figura 9, curiosamente los hombres obtuvieron resultados algo más satisfactorios en su promedio para las 5 diferentes áreas de la prueba. Las mayores diferencias en promedios entre hombres y mujeres se evidenciaron en el área de ciencias naturales y sociales; mientras que el promedio de lectura es muy parecido para ambos sexos, aunque los hombres mantienen una ligera ventaja.

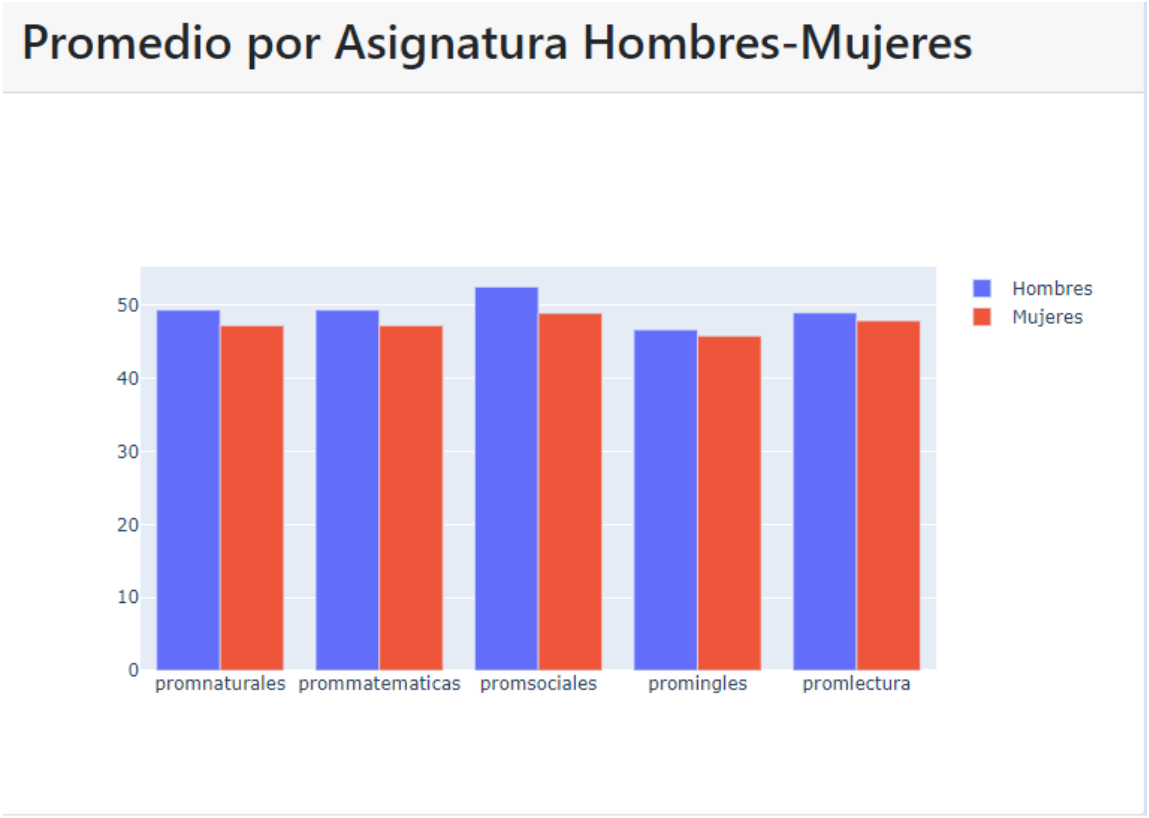


Figura 9: Diagrama de barras sobre promedio del puntaje por asignatura. Hombres y mujeres

4. Discusión de análisis desarrollados.

- **Figura 1 y Figura 2:**
  - **Ventajas:**En la figura 1 y 2 se puede apreciar un entorno agradable y de fácil comprensión para el usuario. Como primera medida, el usuario puede ingresar texto para ubicar de manera más sencilla el departamento al cual quiere estudiar y también puede buscarlo en la barra desplegable. Luego, el usuario observa de manera entendible el promedio del puntaje global obtenido por el colegio perteneciente al departamento que selecciona el ususario, donde fácilmente puede ubicar el mouse sobre una barra correspondiente al nombre de un colegio, y automáticamente se mostrará el número exacto para el promedio del puntaje.
  - **Desventajas:**En algunos casos, el nombre de la institución educativa es muy extenso, por lo que se puede dificultar a primera vista un análisis

completo, sin embargo, existe la opción "zoom in"para hacer la gráfica un poco más grande de tal manera que el usuario pueda comprenderla de manera satisfactoria.

- **Figura 3:**
  - **Ventajas:**La Figura 3 se destaca por su fácil comprensión, organización y brinda la opción de usar un deslizador para clasificar los puntajes según su estrato brindando una fácil interacción con el usuario.
  - **Desventajas:**La gráfica puede llegar a ser monocromática.
- **Figura 4:**
  - **Ventajas:** El diseño de este mapa de calor es estético y comprensible. Refleja de manera clara cuáles son las zonas/departamentos del país que mejor desempeño tienen en cuanto a las pruebas de inglés. La gráfica brinda una noción espacial acerca del análisis realizado.
  - **Desventajas:** No es posible visualizar la región insular de San Andrés y Providencia en el mapa de calor debido a su pequeño tamaño a comparación de las otras regiones.
- **Figura 5:**
  - **Ventajas:**Este diagrama tiene la ventaja de que nos permite visualizar las proporciones de un valor, en este caso el número de estudiantes por estrato. Es de fácil comprensión y es práctico a la hora de inferir los resultados. En este caso, es fácil reflejar la desigualdad social en la educación colombiana. Además, brinda la oportunidad de ver un valor numérico para quién así lo prefiera.
  - **Desventajas:**El grupo no encuentra una desventaja significativa en esta figura.
- **Figura 6:**
  - **Ventajas:** El diagrama de barras es entendible debido a que la caracterización cromática es organizada y diciente. Tiene ventajas similares a la Figura 3.
  - **Desventajas:** Puede existir cierta redundancia entre los colores y la escala.

- **Figura 7:**
  - **Ventajas:**El diagrama de barras horizontal nos permite ver de manera jerárquica y comprensible los puntajes de los mejores diez colegios de Colombia. A su vez, los nombres de los colegios se visualizan de mejor manera que en un diagrama de barras vertical. La escala de puntajes permite visualizar que a pesar de que estos colegios son los mejores del país, existen grandes diferencias entre ellos.
  - **Desventajas:**El diagrama está organizado del menor a mejor puntaje, sin embargo, un aspecto a mejorar sería ubicarlos de manera contaria. También se podría pensar en una gráfica que enumere al dato, en este caso, al colegio.

- **Figura 8:**
  - **Ventajas:** Para el análisis que compete a este diagrama de barras tenemos la ventaja de que únicamente son dos barras a contrastar, por lo que es de fácil comprensión y visualización.
  - **Desventajas:** No encontramos mayores desventajas a la hora de interpretar los resultados o de evaluar su estética.

- **Figura 9:**
  - **Ventajas:** En el diagrama de barras dividido por sexos se destaca la facilidad de comprensión por sus colores atractivos e identificables. A su vez, las barras proporcionan una visualización clara para cada categoría.
  - **Desventajas:** A futuro, se pueden modificar las etiquetas que van debajo de cada área examinada para facilitar la comprensión de los resultados.

5. Conclusiones.

- **Sobre la selección de la fuente de datos:**

  Se utilizaron alrededor de 546000 tuplas (número de pruebas presentadas en el período 2019-II) con los datos para realizar el análisis requerido y así cumplir con los objetivos del proyecto. Estos datos seleccionados son extraídos de una fuente oficial del Estado para así poder garantizar una mayor veracidad en los resultados obtenidos.

Se ha decidido utilizar esta fuente de datos puesto que su múltiple cantidad de columnas permiten la elaboración de diferentes entidades y atributos para un análisis más detallado. Adicionalmente, se aprovecha el fácil acceso a estos recursos de interés público a través del internet por lo cual obtenerlos no fue de mayor dificultad.

- **Sobre el diseño de la base de datos:**

  Se definen 5 entidades (**examen, estudiante, colegio, municipio y departamento**) en nuestro modelo relacional. Estas se disponen sistemáticamente de manera que los diferentes datos esten ordenados para su correcto y eficiente análisis. Cada una de las entidades contiene sus atributos; estos atributos contienen llaves primarias, llaves foráneas y sus respectivas restricciones de acuerdo con las reglas establecidas previamente para el correcto funcionamiento del análisis sobre la base de datos.

  Para la entidad **'examen'** se tienen los atributos **'percentil','puntaje_global'**, **'periodo','puntaje_c_naturales'**, **'puntaje_matematicas','puntaje_sociales'**, **'puntaje_ingles' y 'puntaje_lectura_critica'**. Con **'id_estudiante'** como llave foránea proveniente de la entidad **'estudiente'**. La relación de esta entidad con la entidad **'estudiante'** parten del hecho de que un mismo estudiante puede tomar varios exámenes.

  Para la entidad **'estudiante'** se tienen los atributos **'id_estudiante'**,**'fecha_nacmiento'**, **'estrato' y 'sexo'**. Con **'codigo_colegio'** como llave foránea proveniente de la entidad **'colegio'** y así estableciendo la relación entre el estudiante y su colegio. Esta relación está pensada de manera que un estudiante tiene solo un colegio, pero un colegio sí puede tener muchos estudiantes.

  Para la entidad **'colegio'** se tienen los atributos **'codigo','nombre_colegio'**, **'calendario','genero','localizacion'** y **'naturaleza'**. La llave foránea en este caso es **'codigo_m'** proveniente de la entidad **'municipio'**, estableciendo una relación tal que un municipio puede tener varios colegios y no al revés.

  Para la entidad **'municipio'** se tienen los atributos **'codigo_m', 'nombre_m'**, **'numero_evaluados_m'** y la llave foránea **'codigo_d_departamento'**. Siguiendo la misma línea, un departamento puede tener muchos municipios, pero no viceversa.

  Finalmente para la entidad **'departamento'** se tienen los atributos **'codigo_d','nombre_d'** y **'numero_evaluados_d'**.

  Conclusión del procedimiento: El grupo considera que se logró un diseño de base de datos comprensible y práctico para las consultas que el grupo quiso desarrollar. Además, se establecieron las relaciones necesarias e identificadas para dar cumplimiento a las reglas planteadas en la primera entrega del proyecto. El diseño de la base de datos se encuentra en tercera forma normal, pues

cumple los criterios que estas formas exigen y así la base de datos se vuelve más fácil de manejar para el uso de dash, consultas y cargas de datos. Finalmente, el grupo considera que los atributos son los indicados para identificar y realizar un análisis que puedan reflejar realidades del país en el ámbito de la educación media.

El resultado de este procedimiento se puede reflejar en el punto 2 de esta entrega y en el código de la implementación de la base de datos disponible en el repositorio.

- **Sobre la carga de información:**

  En primer lugar, se realizó la carga para las tablas **'departamento'** y **'municipio'** desde dos archivos .csv disponibles en el repositorio del proyecto.

  En segundo lugar, para continuar con la carga de datos fue necesario revisar nuestra primera implementación de la tabla **'colegio'**. Para los atributos 'codigo' y 'nombre_colegio', se cambió el tipo de dato a varchar(n) y varchar(30) respectivamente. Realizado esto, se cargaron los datos del archivo .csv correspondiente del repositorio.

  En tercer lugar, así como se hizo con la tabla **'colegio'**, también se realizaron algunos cambios a la tabla **'estudiante'**. Para esta carga de datos, fue necesario modificar su correspondiente archivo .csv, pues algunos datos para el atributo 'fecha_nacimiento' presentaban un formato equivocado que no contempla el atributo date de SQL. Esto fue solucionado con el archivo FixingDate.cpp (ubicado en la carpeta code en GitHub) la cual nos adaptó los formatos de este atributo. A su vez, también se realizaron cambios en el atributo estrato, pues algunos estudiantes manifestaban no tener estrato, y a estos datos se le asignó el número -1. Hecho esto, se cargaron los datos desde el archivo .csv

  Finalmente, para la tabla **'examen'**, se removió una restricción 'not null' que había sido implementada al atributo 'puntaje_ingles', esto debido a la falta de algunos datos en el archivo .csv. Se realizó la carga de datos y terminó el procedimiento.

  Conclusión del procedimiento: Durante el proceso de carga de datos el grupo apreció que, a pesar de tener una planificación y diseño de bases de datos frente a una idea original, los diccionarios de datos de gran tamaño muchas veces presentarán inconsistencias, excepciones, errores ortográficos y tipográficos que generan inconvenientes a la hora de cargarlos y de analizarlos. Debido a esto, es necesario que previo al diseño e implementación de la base de datos se analicen los datos lo más que se pueda. Sin embargo, el grupo piensa que la carga de datos fue exitosa y cumplió sus objetivos, los cuales eran distribuir y disponer los datos de tal forma que pudiesen ser consultados y analizados para cumplir el objetivo del proyecto.

- **Sobre la conexión con la base de datos, el desarrollo en dash y los análisis identificados:**

  Durante el procedimiento de conexión con la base de datos no se tuvo ningún inconveniente, además se comprobó el funcionamiento de diferentes consultas para visualizar datos de nuestra base de datos. El grupo considera la conexión con la base de datos como un procedimiento sencillo y muy práctico, pues permite consultar datos desde fuentes externas y realizar análisis con esto.

  Durante la implementación de Dash en Python el grupo se enfrentó distintos retos. Uno de estos retos fue llevar a cabo una idea en la que se planteó implementar una barra despegable en la cual el usuario pudiese interactuar con la aplicación. Seguido a esto, se consultó documentación de dash para llevar a cabo esta idea, la cual se desarrolló exitosamente y se adjuntó a un gráfico de barras en la cual se pueden ver los mejores siete colegios de cada departamento. Otro reto a destacar fue el de implementar el mapa de calor de Colombia, en el cual fue necesario trabajar con archivos .JSON para llevarlo a cabo, finalmente el resultado fue exitoso. Entre otros retos, también se consultó diversa documentación para los decoradores de la aplicación.

  Conclusión del procedimiento: Se logró la meta de crear una aplicación con los distintos análisis identificados en la base de datos. El grupo considera que la aplicación es entendible, manejable e interactiva para los usuarios. Además, los datos son concisos y presentan de forma clara en las gráficas. Por lo tanto, la aplicación es una gran herramienta para analizar los datos del exámen ICFES Saber 11 del periodo 2019-II que a su vez reflejan la situación del país en el ámbito de educación.

6. URL GitHub.

URL del repositorio:

`https://github.com/DavPlazas/Proyecto-Manejo-Bases-de-Datos`

Los archivos correspondientes al código final de la aplicación se encuentra en: `https://github.com/DavPlazas/Proyecto-Manejo-Bases-de-Datos/tree/main/An%C3%A1lisisFinal/CodigoFinal`

Lás imágenes de las gráficas implementadas se encuentran en la carpeta en: `https://github.com/DavPlazas/Proyecto-Manejo-Bases-de-Datos/tree/main/An%C3%A1lisisFinal/Imagenes`

Los archivos de la carga de datos se encuentran en `https://github.com/DavPlazas/Proyecto-Manejo-Bases-de-Datos/tree/main/Archivos%20Carga%20de%20Datos`

GitHub expresaba inconvenientes con el archivo ExamenFianl.csv por su peso. Este archivo está disponible en: `https://uredu-my.sharepoint.com/:x:/g/personal/germand_plazas_urosario_edu_co/EQkYsZ3v4XFKkKbyGtHBIQUBn9isRZFrUPtLPEHwSfZoUw?e=zaTOoJ`