



Enero 30, 2024

Proyecto de Estadística

Juan Antonio González

juangonz@espol.edu.ec

ESPOL (FIEC)

Ramón Ignacio Macías

Ochoa

rimacias@espol.edu.ec

ESPOL (FIEC)

Gary Steven Flores

gsflores@espol.edu.ec

ESPOL (FIMCM)

Germán David Correa

gdcorrea@espol.edu.ec

ESPOL (FIMCM)

Gabriel Cañarte Lucio

lcanarte@espol.edu.ec

ESPOL (FIEC)

Índice

1 Introducción	4
2 Objetivos	5
2.1 Objetivo General	5
2.2 Objetivos Específicos	5
3 Metodología	6
4 Variables de la base de Datos	7
4.1 Variables Cuantitativas	7
4.1.1 Estudiantes Blancos, Negros y Asiáticos	7
4.1.2 Promedios de SAT	8
4.1.3 Estudiantes Testados	9
4.2 Variables Cualitativas	10
4.2.1 Borough	10
4.2.2 City	12
5 Análisis y Resultados	13
5.1 Análisis de Correlación entre el promedio SAT y etnias	13
5.2 ANOVA ONE WAY entre variables promedio SAT y City	15
5.3 Exploración de la Relación entre el tamaño de la población estudiantil y las puntuaciones SAT mediante regresión lineal	16
6 Conclusiones	20
6.1 Análisis de varianza entre promedios SAT y ubicación	20
6.2 Correlación entre la diversidad étnica y las puntuaciones SAT	21
6.3 Exploración de la Relación entre el tamaño de la población estudiantil y las puntuaciones SAT mediante regresión lineal	22
7 Anexos	23
7.1 Código Fuente	23
Bibliografía	24

Índice de Figuras

Figura 1: Diagrama de cajas de los porcentajes de estudiantes blancos, negros y asiáticos.	7
Figura 2: Diagrama de cajas de los promedios de la masterias SAT.	8
Figura 3: Histograma de estudiantes testados	9
Figura 4: Gráfico de pastel de la variable cualitativa Borough	10
Figura 5: Gráfico de barras de frecuencias de la variable cualitativa City.	12
Figura 6: Gráfico de correlación #1	13
Figura 7: Gráfico de correlación #2	13
Figura 8: Resultados del ANOVA ONE WAY entre variables promedio SAT y City	15
Figura 9: Gráfico de regresión lineal - Math	16
Figura 10: Gráfico de regresión lineal - Reading	17
Figura 11: Gráfico de regresión lineal - Writing	18

1 Introducción

El presente informe se enfoca en el análisis de las puntuaciones SAT en las escuelas del estado de Nueva York. La elección de este conjunto de datos se basa en la relevancia y el impacto que tienen estas puntuaciones en la educación y el futuro de los estudiantes en la ciudad de Nueva York, Estados Unidos. El análisis de estas puntuaciones puede proporcionar información valiosa sobre la calidad de la educación en diferentes escuelas y distritos, lo que puede ser útil para los responsables de la toma de decisiones en el ámbito de la educación.

La base de datos a analizar incluye información detallada sobre las escuelas de Nueva York, el nombre de la escuela, el distrito, ciudad, estado, horario de inicio y fin, matrícula de estudiantes y porcentaje de estudiantes de diferentes grupos étnicos. Además, se proporcionan las puntuaciones promedio de SAT en asignaturas como matemáticas, lectura y escritura, así como el porcentaje de estudiantes que realizaron la prueba.

2 Objetivos

2.1 Objetivo General

Evaluar las puntuaciones SAT en las escuelas de Nueva York y determinar si existen diferencias significativas en las puntuaciones en función de la ubicación geográfica y la diversidad étnica.

2.2 Objetivos Específicos

1. Identificar diferencias en las puntuaciones SAT en función de la ubicación geográfica
2. Investigar la correlación entre la diversidad étnica y las puntuaciones SAT
3. Encontrar la relación entre el tamaño de la población estudiantil y las puntuaciones SAT utilizando regresión lineal

3 Metodología

Para la realización de este proyecto se utilizó el lenguaje de programación **R**, empleando un conjunto de librerías que facilitaron el análisis de los datos.

De la base de datos original se extrajeron las variables que se consideraron de valor para el análisis, las cuales fueron:

Variables cualitativas:

- Borough: Distrito de la escuela
- City : Ciudad de la escuela

Variables cuantitativas:

- Student_Enrollment: Cantidad de estudiantes matriculados
- Percent_White: Porcentaje de estudiantes blancos
- Percent_Black: Porcentaje de estudiantes negros
- Percent_Asian: Porcentaje de estudiantes asiáticos
- Average_Score_SAT_Math: Promedio de las puntuaciones SAT en matemáticas
- Average_Score_SAT_Reading: Promedio de las puntuaciones SAT en escritura
- Average_Score_SAT_Writing: Promedio de las puntuaciones SAT en lectura
- Percent_Testes: Porcentaje de estudiantes que realizaron la prueba
- Average_SAT: Promedio de las puntuaciones SAT

4 Variables de la base de Datos

4.1 Variables Cuantitativas

4.1.1 Estudiantes Blancos, Negros y Asiáticos

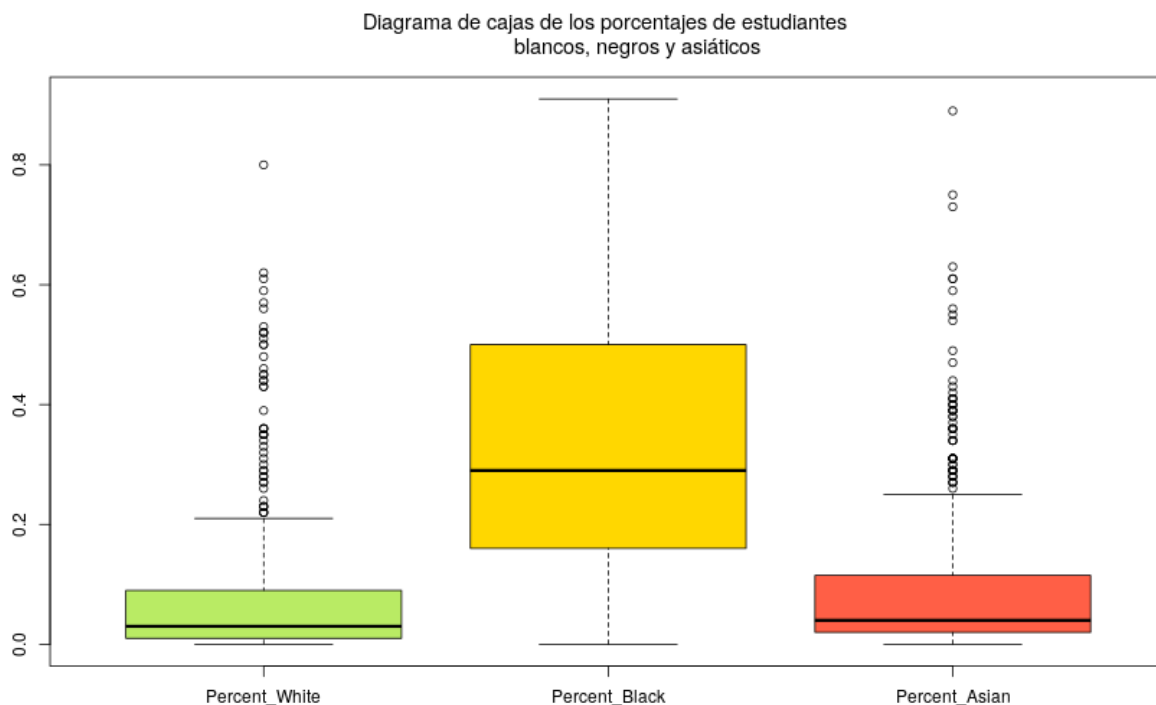


Figura 1: Diagrama de cajas de los porcentajes de estudiantes blancos, negros y asiáticos.

El presente gráfico de cajas representa los porcentajes de estudiantes de etnias blancos, negros y asiáticos. Los estudiantes blancos, representados por la caja verde, tienen un rango de porcentajes que varía entre 0.01 y 0.61, con una mediana alrededor de 0.04 y algunos valores atípicos en el extremo superior. Los estudiantes de etnia negros, representados por la caja amarilla, tienen un rango de porcentajes que varía entre 0.03 y 0.53, con una mediana alrededor de 0.28 y algunos valores atípicos en ambos extremos. Los estudiantes asiáticos, representados por la caja roja, tienen un rango de porcentajes que varía entre 0.0 y 0.89, con una mediana alrededor de 0.09 y varios valores atípicos en el extremo superior.

4.1.2 Promedios de SAT

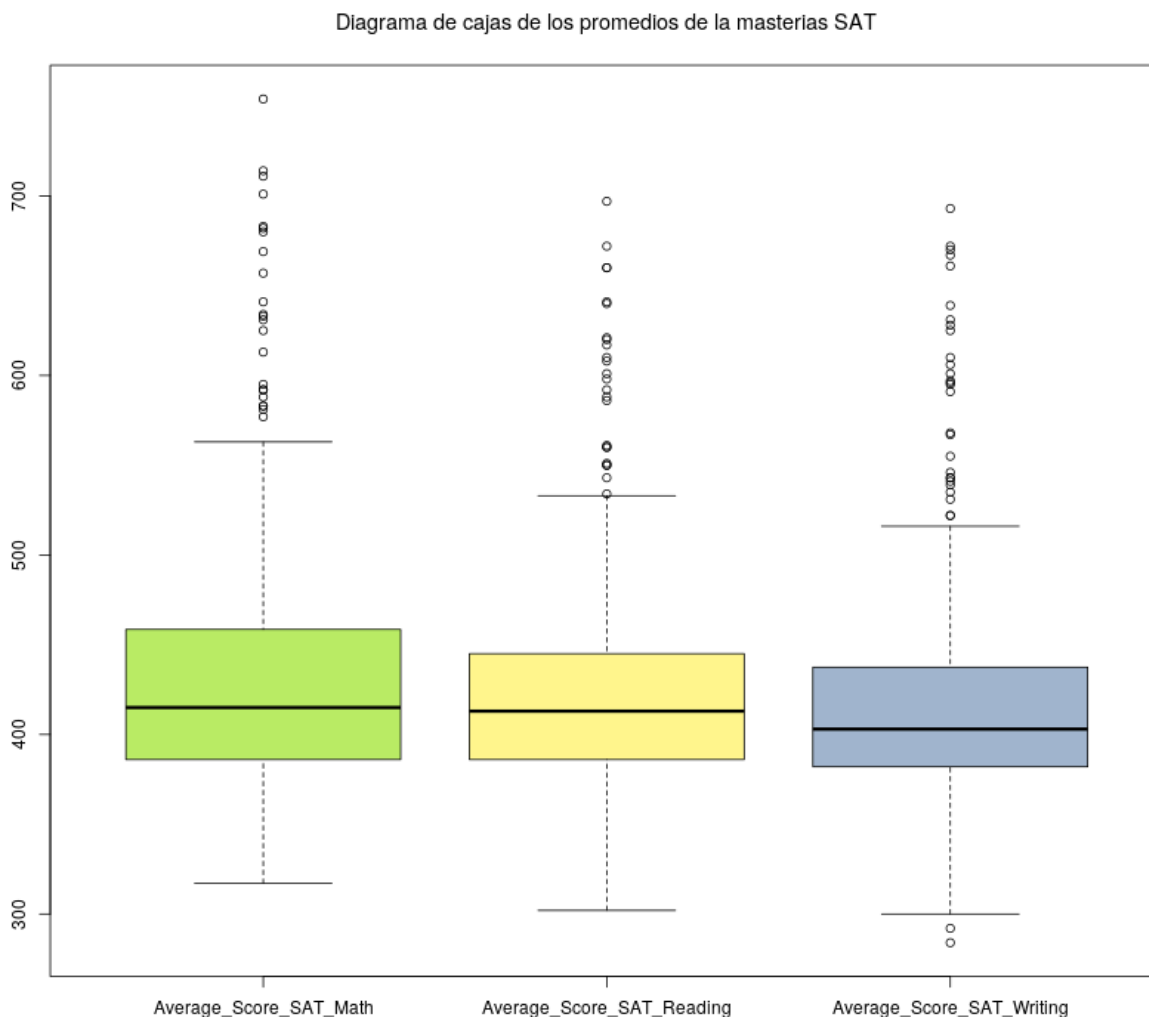


Figura 2: Diagrama de cajas de los promedios de la masterías SAT.

El presente gráfico de cajas que representa los promedios de las puntuaciones SAT en Matemáticas, Lectura y Escritura. En Matemáticas, el 25% de los estudiantes obtuvo una puntuación de 386 o menos, el 50% obtuvo una puntuación de 415 o menos, y el 75% obtuvo una puntuación de 459 o menos. En Lectura, los cuartiles son 386, 413 y 445 respectivamente, lo que indica que las puntuaciones en lectura son generalmente más altas que en matemáticas. En Escritura, los valores son 382, 403 y 438 para cada uno de los tres cuartiles respectivamente, lo que sugiere que las puntuaciones en escritura son generalmente más bajas que en las otras dos asignaturas. Los puntos por encima de cada caja representan posibles valores atípicos, es decir, puntuaciones que se desvían significativamente del resto de los datos.

4.1.3 Estudiantes Testados

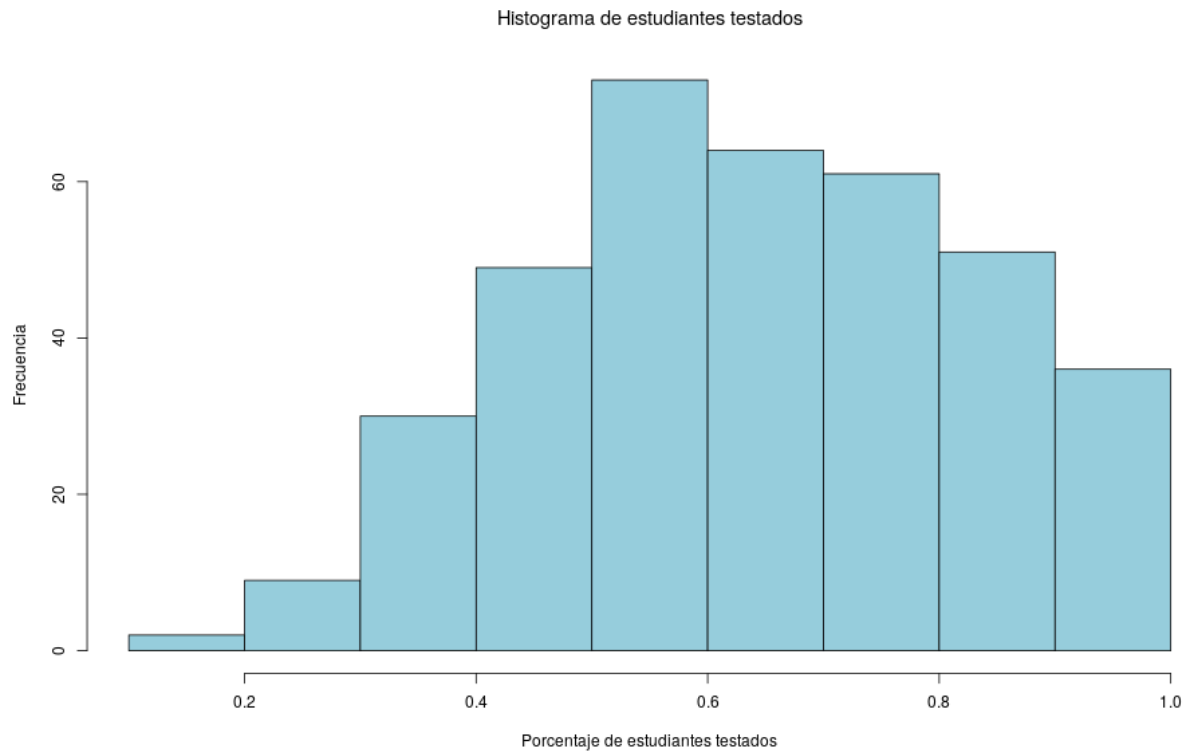


Figura 3: Histograma de estudiantes testados

La gráfica de estilo histograma anterior representa la distribución de los porcentajes de estudiantes que realizaron la prueba **SAT**. La mayor frecuencia de estudiantes testados se encuentra en el rango de 0.5 a 0.7. Con respecto a la distribución de los datos, se encuentra una distribución no uniforme, con un comportamiento irregular.

4.2 Variables Cualitativas

4.2.1 Borough

Gráfico de pastel de la variable cualitativa Borough

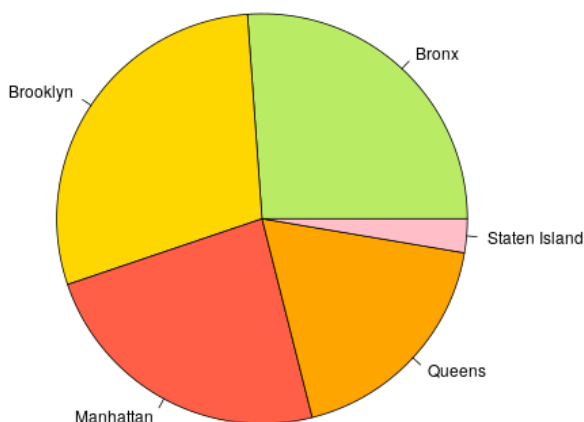


Figura 4: Gráfico de pastel de la variable cualitativa Borough

El gráfico de pastel presentado muestra una representación visual detallada de la distribución porcentual de la población en los cinco distritos de la ciudad de Nueva York, que son Brooklyn, Bronx, Manhattan, Queens y Staten

Island. Cada segmento del gráfico de pastel representa un distrito y su tamaño es proporcional al porcentaje de la población que reside en ese distrito. Brooklyn, con el 29,07% de la población, tiene la mayor proporción, lo que se refleja en el segmento más grande del gráfico de pastel. Esto indica que Brooklyn es el distrito más poblado de la ciudad de Nueva York. El Bronx, con el 26,13% de la población, tiene el segundo segmento más grande del gráfico de pastel, lo que indica que también es un distrito densamente poblado. Manhattan, tiene el tercer segmento más grande en el gráfico de pastel, representando el 23,73% de la población. Queens, tiene el cuarto segmento más grande en el gráfico de pastel, representando el 18,40% de la población. Staten Island, tiene el segmento más pequeño en el gráfico de pastel, representando solo el 2,67% de la población. Esto indica que Staten Island tiene la

menor densidad de población entre los cinco distritos. La suma total de las proporciones es del 100,00%, lo que indica que todos los distritos de la ciudad de Nueva York están representados en este gráfico.

4.2.2 City

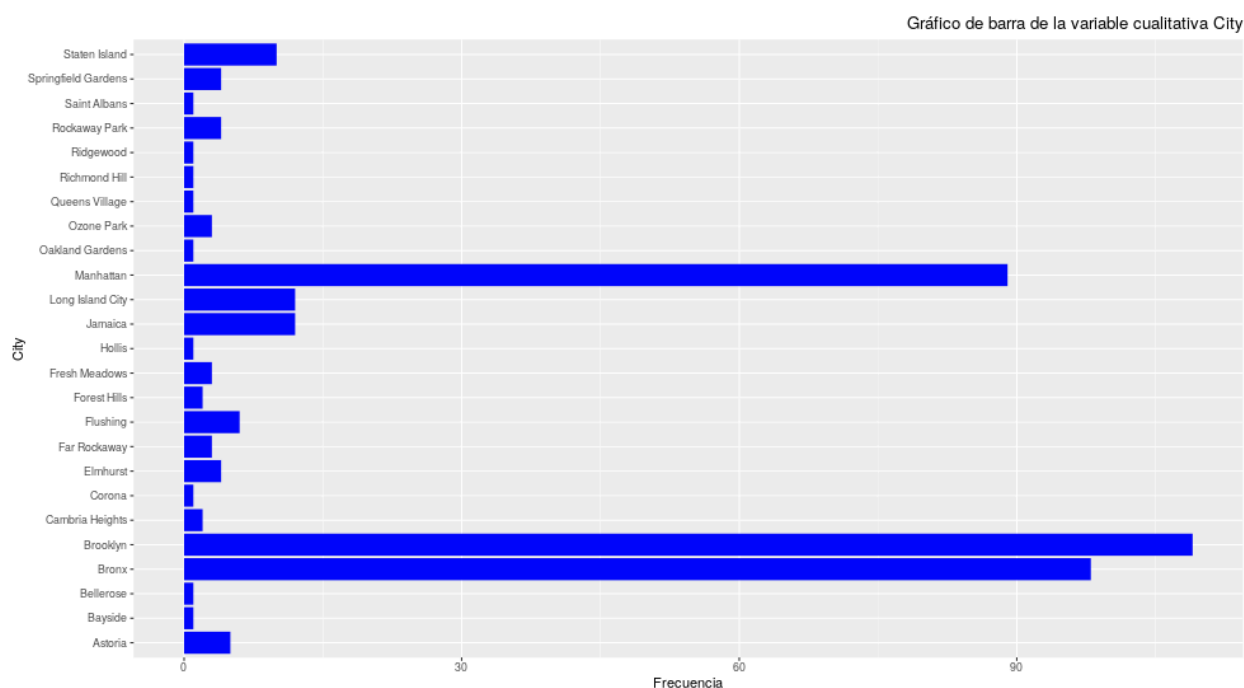


Figura 5: Gráfico de barras de frecuencias de la variable cualitativa City.

El presente diagrama barras presenta la distribución de las ciudades en donde habitan los estudiantes de la base estudiada, los datos mostraron que Brooklyn encabeza la lista con un total de 109 ciudades. Esto indica que Brooklyn es una región con una gran cantidad de ciudades en su territorio. Le sigue de cerca el Bronx, con 98 ciudades, y Manhattan, con 89 ciudades. Por otro lado, Staten Island alberga 10 ciudades, lo que es significativamente menor en comparación con Brooklyn, Bronx y Manhattan. Esto podría indicar que Staten Island tiene una densidad de ciudades mucho menor. Además, hay varios lugares como Bayside, Bellerose, Corona, Hollis, Oakland Gardens, Queens Village, Richmond Hill, Ridgewood y Saint Albans que tienen solo una ciudad cada uno. Esto podría sugerir que estas áreas son menos urbanizadas o que están compuestas principalmente por una sola ciudad grande.

5 Análisis y Resultados

5.1 Análisis de Correlación entre el promedio SAT y etnias

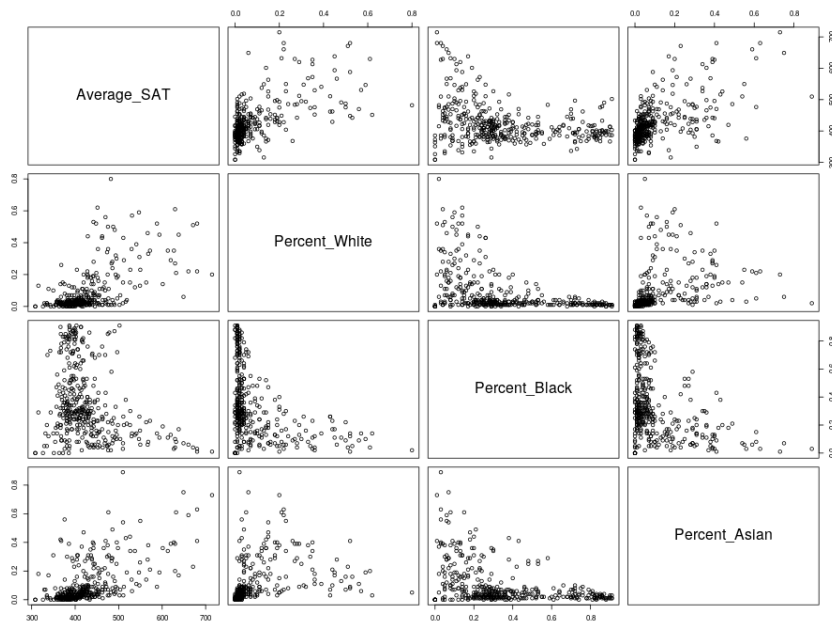


Figura 6: Gráfico de correlación #1

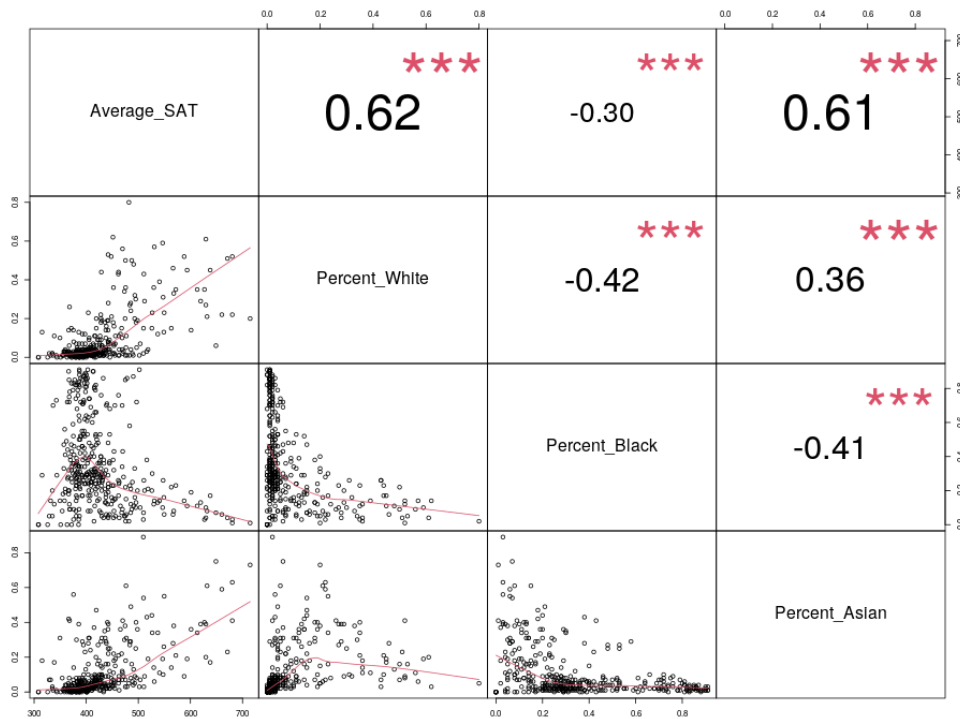


Figura 7: Gráfico de correlación #2

En la Figura 7 se presentan los resultados de la correlación entre las variables Average_SAT, Percent_White, Percent_Black y Percent_Asian. Se puede observar que la correlación entre Average_SAT y Percent_White es positiva con un valor de 0.6234902, lo que indica que existe una relación directa entre estas dos variables. Por otro lado, la correlación entre Average_SAT y Percent_Black es negativa con un valor de -0.3048109, lo que sugiere una relación inversa entre ambas variables. Así mismo, se puede apreciar que la correlación entre Average_SAT y Percent_Asian es positiva con un valor de 0.6098355, lo que indica una relación directa entre estas dos variables. Por último, la correlación entre Percent_White y Percent_Black es negativa con un valor de -0.4220592, mientras que la correlación entre Percent_White y Percent_Asian es positiva con un valor de 0.3555783.

Estos resultados muestran que existe una relación entre las variables analizadas y que es importante tener en cuenta su impacto en los resultados del **SAT**.

5.2 ANOVA ONE WAY entre variables promedio SAT y City

```
      Df Sum Sq Mean Sq F value    Pr(>F)
data$City  24  258367    10765    2.854 1.45e-05 ***
Residuals 350 1320272     3772
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 8: Resultados del ANOVA ONE WAY entre variables promedio SAT y City

La figura anterior muestra el análisis de los resultados del ANOVA en donde se revela una diferencia significativa entre las variables Average_SAT y City.

Según los datos obtenidos, se puede concluir que la ciudad de residencia de los estudiantes influye de manera significativa en el rendimiento promedio del examen **SAT**. Con un valor de F de 2.854 y un p-valor de 1.45e-05, se puede afirmar con seguridad que existe una relación entre ambas variables. Pese a esto, es importante tener en cuenta que aún hay un alto porcentaje de variabilidad que no puede ser explicado por la ciudad de residencia, lo cual sugiere que existen otros factores que también pueden influir en el rendimiento en el examen SAT.

5.3 Exploración de la Relación entre el tamaño de la población estudiantil y las puntuaciones SAT mediante regresión lineal

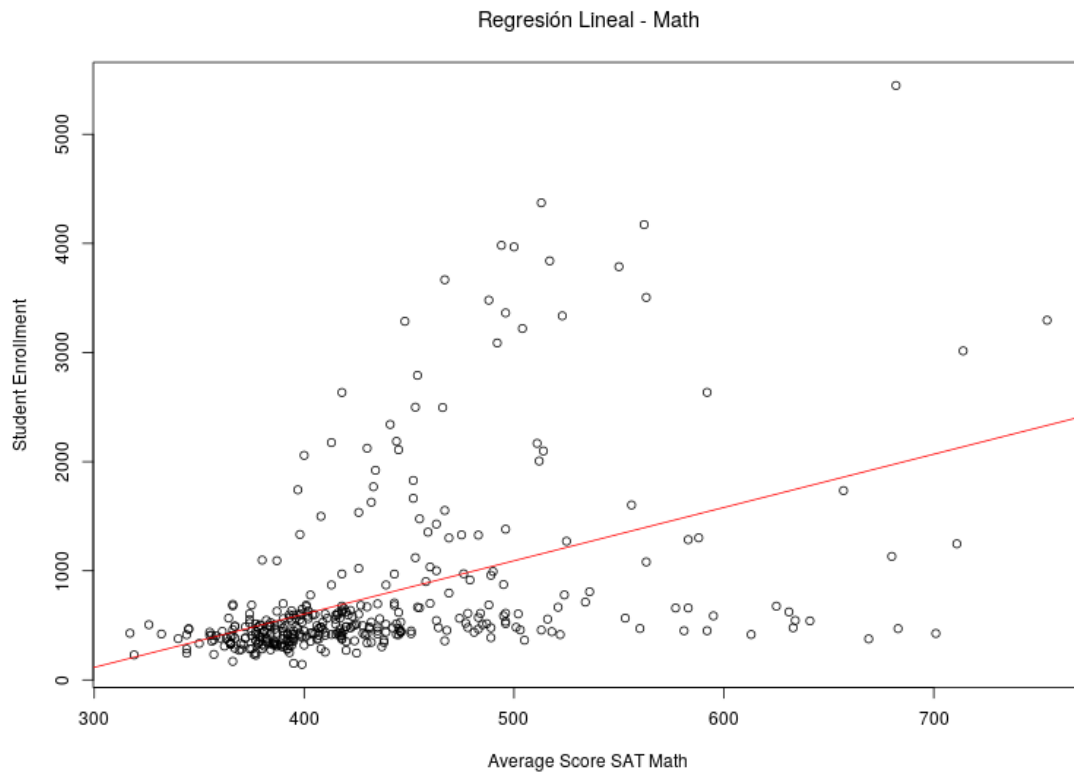


Figura 9: Gráfico de regresión lineal - Math

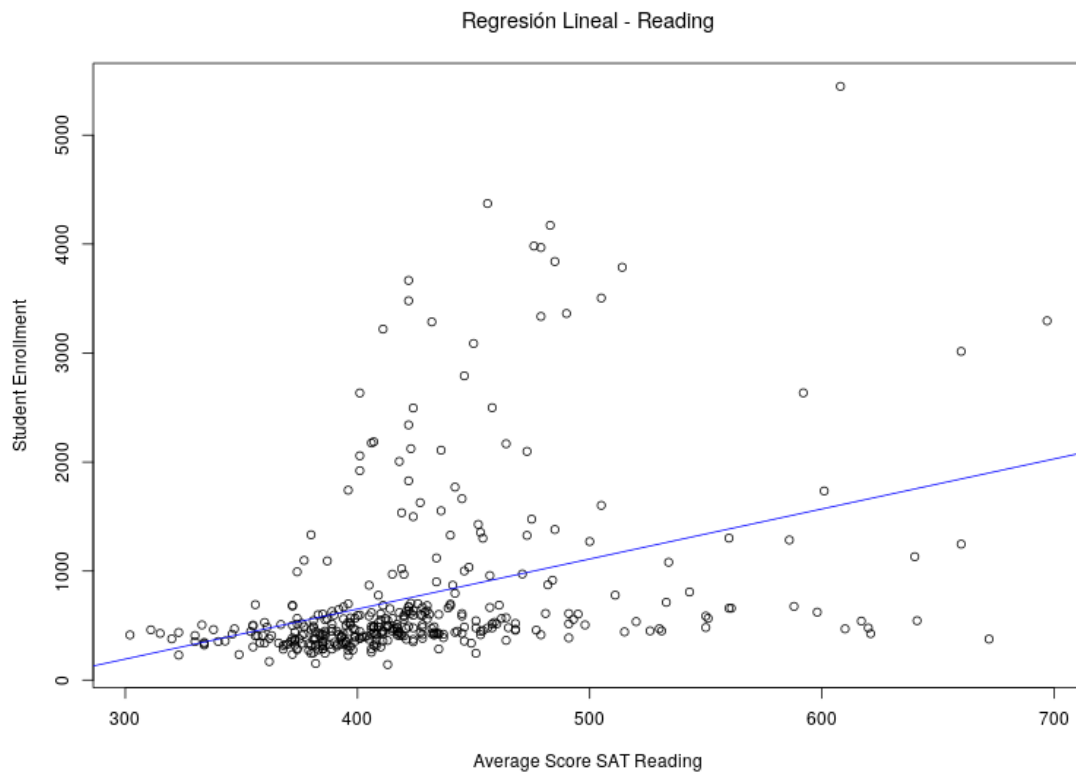


Figura 10: Gráfico de regresión lineal - Reading

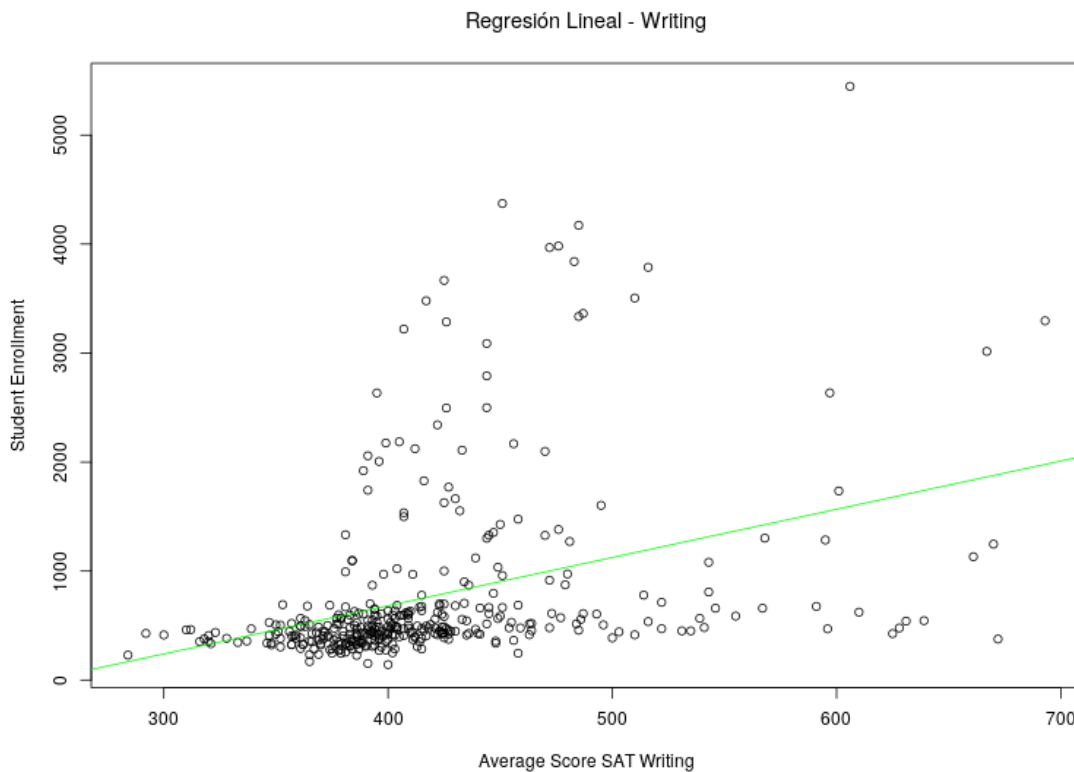


Figura 11: Gráfico de regresión lineal - Writing

El objetivo es aplicar técnicas de regresión lineal para examinar la relación entre el tamaño de la matrícula estudiantil y las puntuaciones promedio en los exámenes SAT de matemáticas, lectura y escritura. La regresión lineal proporciona un enfoque sistemático para modelar la dependencia entre estas variables, permitiendo obtener coeficientes clave como la pendiente e intercepto.

A través de cálculos y análisis realizados en el entorno estadístico R, se determinarán los parámetros fundamentales de cada modelo de regresión lineal. Estos parámetros se traducirán en ecuaciones lineales específicas que describirán la relación cuantitativa entre el tamaño de la matrícula y el desempeño académico en diferentes áreas.

Las rectas obtenidas son las siguientes:

Para la puntuación promedio de Matemáticas se tiene $y = 4.88x - 1349$.

Para la puntuación promedio de Lectura se tiene $y = 4.59x - 1183.83$.

Para la puntuación promedio de Escritura se tiene $y = 4.424x - 1086.576$.

Al evaluar las ecuaciones de regresión lineal para las puntuaciones promedio de Matemáticas, Lectura y Escritura en relación con el tamaño de la matrícula estudiantil, observamos una tendencia inesperada pero consistente: todas muestran una relación positiva. En el caso de Matemáticas, la pendiente positiva de $y = 4.88x - 1349$ sugiere que, a medida que aumenta la matrícula, las puntuaciones promedio de Matemáticas tienden a mejorar. Esta dinámica también se refleja en las otras dos áreas evaluadas, donde tanto Lectura ($y = 4.59x - 1183.83$) como Escritura ($y = 4.424x - 1086.576$) muestran la misma inclinación positiva. Más allá de la sorprendente tendencia positiva con respecto al tamaño de la matrícula, es crucial considerar otros factores que podrían estar influyendo en estas relaciones. Elementos como la calidad de la enseñanza, recursos educativos disponibles y programas de apoyo podrían estar contribuyendo al rendimiento académico en formas complejas. Además, se podría explorar la posibilidad de variaciones regionales o socioeconómicas que podrían impactar la relación entre el tamaño de la población estudiantil y el éxito académico.

6 Conclusiones

6.1 Análisis de varianza entre promedios SAT y ubicación

En base a los hallazgos derivados del análisis ANOVA, se puede concluir que la ubicación geográfica de las escuelas en Nueva York tiene una influencia significativa en el rendimiento promedio de los estudiantes en el examen SAT. El estudio revela diferencias estadísticamente significativas entre las puntuaciones SAT y la ciudad de residencia de los estudiantes, señalando que factores asociados con la localización geográfica desempeñan un papel crucial en las disparidades observadas. Esta conclusión sugiere que no todas las ubicaciones geográficas ofrecen el mismo entorno educativo, y factores contextuales relacionados con la ciudad de residencia podrían estar contribuyendo de manera destacada a las variaciones en el rendimiento académico.

A pesar de la evidente influencia de la ubicación geográfica, es esencial reconocer que existe un alto porcentaje de variabilidad en las puntuaciones SAT que no puede ser completamente explicado por este factor. Esta realidad subraya la complejidad del fenómeno y apunta hacia la existencia de otros factores no abordados en el estudio que también desempeñan un papel determinante en las diferencias de rendimiento. Para comprender plenamente estas variaciones, futuras investigaciones podrían dirigirse hacia la identificación de estos factores adicionales y explorar su contribución al rendimiento en el examen SAT.

En última instancia, el reconocimiento de disparidades en las puntuaciones SAT basadas en la ubicación geográfica subraya la necesidad de adoptar un enfoque integral al enfrentar diferencias en el desempeño académico. Este planteamiento va más allá de la educación per se y considera los contextos particulares de las ciudades, examinando su impacto en la accesibilidad a recursos educativos, la excelencia docente y otros factores socioeconómicos que pueden incidir en el éxito estudiantil.

6.2 Correlación entre la diversidad étnica y las puntuaciones SAT

Ahora bien, en la exhaustiva investigación sobre la correlación entre la diversidad étnica y las puntuaciones SAT en el contexto de las escuelas de Nueva York, se concluye que existen patrones notables en el rendimiento académico asociados a la composición étnica de los estudiantes. Los resultados revelan una correlación positiva entre las puntuaciones SAT y los porcentajes de estudiantes blancos y asiáticos, indicando que estos grupos étnicos tienden a obtener puntuaciones más altas en el examen. En contraste, se observa una correlación negativa con el porcentaje de estudiantes negros, sugiriendo que este grupo enfrenta desafíos adicionales que pueden reflejarse en puntuaciones más bajas.

Aunque se observe claridad en estas conexiones, es crucial considerar que estas relaciones no son predecibles y pueden ser influidas por diversos factores del entorno. Este estudio subraya la necesidad de un análisis más detenido para comprender la complejidad de las disparidades en la educación y el efecto de la diversidad étnica en el desempeño académico. La exploración de elementos adicionales, como la calidad de la enseñanza, la disponibilidad de recursos y las condiciones socioeconómicas, puede brindar una perspectiva más exhaustiva de la relación entre diversidad étnica y logros académicos.

En definitiva, la detección de estas relaciones indica la relevancia de aplicar tácticas educativas adaptadas para atender las particularidades de cada comunidad étnica. Asimismo, resalta la urgencia de políticas educativas inclusivas que impulsen la igualdad y solventen las desigualdades identificadas, propiciando un entorno académico que respalde el éxito de todos los estudiantes, sin importar su origen étnico.

6.3 Exploración de la Relación entre el tamaño de la población estudiantil y las puntuaciones SAT mediante regresión lineal

Con respecto a los análisis basados en técnicas de regresión lineal revelan una relación positiva entre el tamaño de la matrícula estudiantil y las puntuaciones promedio en los exámenes SAT de matemáticas, lectura y escritura. Esta tendencia sugiere que, en promedio, las escuelas con matrículas más grandes pueden experimentar un rendimiento académico mejorado en estas áreas.

Estos resultados reflejan un panorama educativo en el que las escuelas con matrículas más grandes podrían estar potencialmente en una posición ventajosa, ya que parece haber una correlación entre un mayor número de estudiantes y un mejor desempeño académico en múltiples áreas evaluadas.

Aun así, mientras estos hallazgos apuntan hacia una relación positiva entre el tamaño de la matrícula y el rendimiento académico, es importante ejercer cautela en la interpretación de estos resultados. Si bien los datos muestran una tendencia general, no deben pasar por alto las variaciones individuales y los posibles factores de confusión.

7 Anexos

7.1 Código Fuente

El código fuente de este proyecto se encuentra disponible en el siguiente enlace a GitHub

`<https://github.com/anntnrb/estg1034-proy>.`

Bibliografía

- Arce, R., Fariña, F., & Sobral, J. (1995). An approach to statistical prediction of biases in judgment-making by jurors. *International Journal of Social Psychology*, 10(1), 3-16. <https://doi.org/10.1174/021347495763835229>
- Benesty, J., Chen, J., & Huang, Y. (2009). *Correlation and dependence*. Springer.
- Girden, E. R. (1992). *ANOVA: Repeated measures* (Número 84). Sage.
- Manuel, J., & Fernández, V. (1991). Consistencia de un estimador no parametrico, recursivo, de la regresion bajo condiciones generales. *TDE*, 6, 11-31. <https://doi.org/10.1007/BF02873520>
- R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. <https://www.r-project.org/>
- Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis* (Vol. 936). John Wiley & Sons.
- Álvarez-Olguín, G., & Escalante-Sandoval, C. A. (2016). Análisis de frecuencias no estacionario de series de lluvia anual. *Tecnología y ciencias de agua*, 7(1), 71-88.