Tipología y ciclo de vida de los datos

PRA2: ¿Cómo realizar la limpieza y análisis de datos?

Presentado por: Jhon Jairo Realpe

1. Descripción del dataset

El conjunto de datos de estudio contiene información sobre el promedio de calificaciones (GPA) y Scholastic Assessment Test SAT, de los estudiantes matriculados en una universidad de Estados Unidos (ver tabla 1). El número de observaciones es igual a 4137 test asociados al SAT, que es el examen nacional estandarizado que realizan todos los estudiantes universitarios cada cuatro años (Econpapers, 2000).

Tabla 1. Descripción del conjunto de datos [fuente: propia]

No	Variable	Descripción	Tipo
1	sat	Examen SAT combinado (lectura crítica/escritura)	Continua
2	tothrs	Total de horas cursadas de clase hasta terminar el semestre	Continua
3	colgpa	Promedio de notas al terminar el semestre	Continua
4	athlete	= 1 si es atleta	Categórica
5	verbmath	competencia verbal/competencia matemática del examen SAT	Continua
6	hsize	Tamaño del total de graduados en escuelas secundaria(en cientos)	Continua
7	hsrank	rango respecto al total de graduados	Continua
8	hsperc	Percentil del total de graduados respecto al total de estudiantes	Continua
9	female	= 1 si es mujer	Categórica
10	white	= 1 si es blanco	Categórica
11	black	= 1 si es negro	Categórica
12	hsizesq	Total de graduados en escuelas secundaria(al cuadrado)	Continua

En la tabla 1 se aprecia que el conjunto de datos tiene 12 variables, donde athelete, female, white y black son de tipo categórico y las variables restantes son continuas.

1.1. ¿Por qué es importante y qué pregunta/problema pretende responder?

Los estudios relacionados con el rendimiento escolar de los estudiantes a nivel de secundaria, basados en técnicas estadísticas o de machine learning, pueden brindar información valiosa a la institución donde se desarrolla el estudio. Por ejemplo, en función de algunas variables, propias de la dinámica académica / institucional, o de características propias de los estudiantes, tal como etnia, raza, nivel socioeconómico, o también aspectos sociales, políticos y económicos del lugar donde se ubica la institución, se puede tratar de identificar qué factores son los más relevantes y que contribuyen tanto positiva como negativamente en el desempeño escolar de los estudiantes. Identificados estos factores, se pueden proponer estrategias, planes o políticas, orientados a mitigar estas causas, ya sea desde la perspectiva de la institución educativa, o desde una visión más general, a través de políticas educativas desarrolladas por las instituciones gubernamentales que gestionan el sistema de educativo.

2. Limpieza de los datos

2.1. ¿Los datos contienen ceros o elementos vacíos?

Se verifica que el conjunto de datos no tiene ni valores vacíos ni nulos, por ende, no se aplica ninguna técnica correctiva (ver tabla 2)

Tabla 2. Conteo de valores nulos y vacíos en el conjunto de datos [fuente propia]

variable	valores nulos	valores vacíos
sat	0	0
tothrs	0	0
colgpa	0	0
athlete	0	0
verbmath	0	0
hsize	0	0
hsrank	0	0
hsperc	0	0
female	0	0
white	0	0
black	0	0
hsizesq	0	0

2.2. Identifica y gestiona los valores extremos

Para cada atributo numérico se grafica un histograma y un diagrama de caja para identificar valores extremos (ver figura 1).

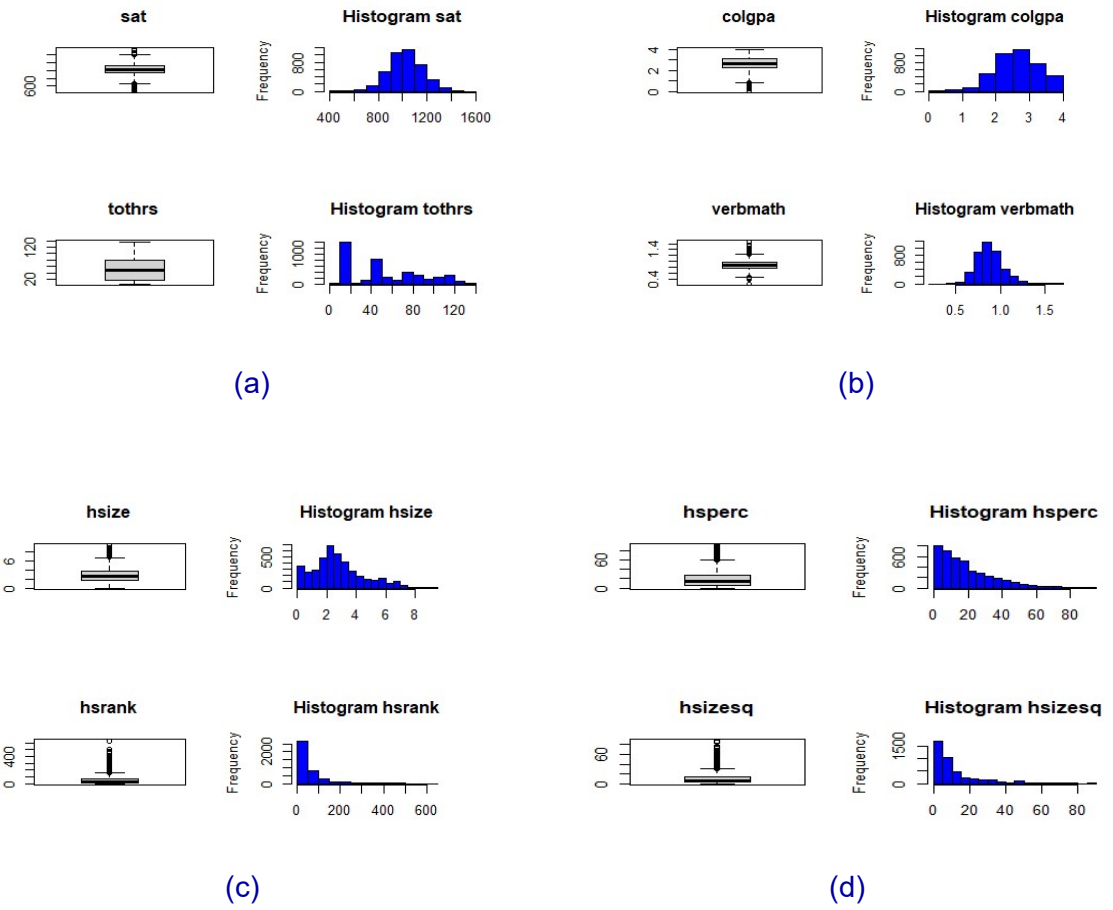

Figura 1. Histograma y gráfico de bigotes para variables continuas [fuente propia]

En la figura se observa que las distribuciones de hsrank, hsize, hsperc y hisizesq, tiene un sesgo a la derecha y una considerable cantidad de valores extremos.

Con la función stats del comando boxplot, se extraen el número de valores extremos, es decir, todo valor que está fuera de los bigotes, que son las líneas que se determinan como el tercer cuartil +1.5 veces el rango intercuartílico y el primer cuartil -1.5 veces el rango intercuartílico.

A continuación, se hace un resumen de los valores extremos en las variables numéricas.

Tabla 3. Conteo de valores extremos para variables continuas [fuente: propia]

Variable	Total valores extremos
hsizesq	415
hsrank	267
hsize	169
hsperc	135
verbmath	82
sat	39
colgpa	34

tothrs	0

Con en esta información, se exponen los criterios para conservar o eliminar los valores extremos.

En primer lugar, las observaciones de la variable hsizesq, presentan un elevado número de valores extremos, esto se debe a que dicha variable es el valor cuadrado de la variable hsize, y por tanto los valores se hacen más grandes y la distribución se sesga hacia la derecha. Dado que esta transformación no mejora las propiedades estadísticas de hsize, no se toma ninguna acción respecto a los valores atípicos y se prescindirá de ella, en la sección de resolución de problemas y el análisis de modelos predictivos.

Respecto a la variable hsrank un 6.4% de los datos son valores extremos, para la variable hsize un 4.1% y para la variable hsperc 3.2%. Aunque la proporción es considerable y hace que las distribuciones tengan un elevado sesgo a la derecha, no se tiene información adicional del proceso de muestreo del experimento realizado. Por tal razón, no se tiene evidencia para descartar los valores y se decide no tomar ninguna acción correctiva.

Con el propósito de abordar el problema del sesgo en las distribuciones, en la sección 5 del presente documento, se procederá a aplicar la transformación más adecuada, de modo que las variables tiendan a una distribución normal.

Respecto a la variable verbmath, los valores se obtienen de la división, entre el examen de competencia verbal SAT, cuyo rango es de 200-800 y el examen de competencia matemática SAT cuyo rango es de 200-800, tal como se reporta en (Forstall, 2019; Muniz, 2021).

Dicho lo anterior, y de acuerdo a la gráfica, su distribución estaría centrada alrededor de 1, además presenta un comportamiento normal y por tal razón, se consideran datos válidos y no se toma ninguna acción correctiva.

Respecto a las variables colgpa y sat, en primer lugar, se corrobora que los rangos sean coherentes en relación a los estándares definidos para estas variables. En este sentido los valores de colgpa están en el rango de 0-4 y para sat de 400-1600, tal como se reporta en (College Board Sat Program, 2022; Muniz, 2021; PrepScholar, 2018; Zhang, 2018)

Con base en esto y de acuerdo a los gráficos anteriores, los valores extremos están en los rangos estándar y la presencia minoritaria de valores extremos, puede deberse a estudiantes con buen desempeño académico. En tal sentido, se consideran datos válidos y no se tomará ninguna acción correctiva.

Respecto a la variable tothrs no realiza ninguna acción ya que no presenta valores extremos

3. Análisis de los datos

En esta sección se realiza el análisis de los datos, el cual se dividirá en dos apartados. En el primero se desarrolla el análisis de las variables continuas y en el segundo el análisis de las variables categóricas.

3.1. Análisis de variables continuas

3.1.1. Análisis descriptivo

En primer lugar, se realiza un análisis descriptivo de las variables numéricas, donde se presentan medidas de tendencia central.

Tabla 4. Medidas de tendencia central [fuente: propia]

	sat	tothrs	colgpa	verbmath	hsize	hsrank	hsperc	hsizesq
Min.	470	6	0.000	0.2597	0.03	1	0.1667	0.0009
1st Qu.	940	17	2.210	0.7759	1.65	11	6.4328	2.7225
Median	1030	47	2.660	0.8667	2.51	30	14.5833	6.3001
Mean	1030	52.83	2.653	0.8805	2.8	52.83	19.2371	10.8535
3rd Qu.	1120	80	3.120	0.9649	3.68	70	27.7108	13.5424
Max.	1540	137.0	4.000	1.6667	9.4	634.00	92.0000	88.36

Para complementar el análisis, se aplican otros estadísticos denominados robustos, ya que no se ven influenciados por valores extremos.

Tabla 5. Medidas de tendencia central robustas [fuente: propia]

	sat	**tothrs**	**colgpa**	**verbmath**	**hsize**	**hsrank**	**hsperc**	**hsizesq**
mean_num	1.03E+03	52.8322	2.6527	0.8805	2.7997	52.8301	19.2371	10.8534
median_num	1.03E+03	47.0000	2.6600	0.8667	2.5100	30.0000	14.5833	6.3001
var_num	1.94E+04	1248.180	0.4338	0.0222	3.0157	4183.9660	274.5227	159.3415
sd_num	1.39E+02	35.3296	0.6586	0.1491	1.7366	64.6836	16.5687	12.6231
skewness_num	7.33E-02	0.4963	-0.2146	0.6496	0.7527	2.5719	1.2914	1.9817
kurtosis_num	1.84E-01	-1.0002	0.1107	1.0672	0.2676	9.1379	1.5286	4.0992
mean_trim	1.03E+03	51.2706	2.6632	0.8750	2.7118	43.9938	17.7336	9.2778
mean_win	1.03E+03	52.6379	2.6654	0.8784	2.7708	48.7845	18.6386	10.3589
IQR_num	1.80E+02	63.0000	0.9100	0.1890	2.0300	59.0000	21.2781	10.8199
mad_num	1.33E+02	45.9606	0.6672	0.1407	1.4233	35.5824	14.0459	6.6385

Para las variables sat, colgpa y verbmath, los valores de la media normal y la media winsorizada (mean_win) y media trim, son similares, lo cual es coherente dado que dichas variables tienen una distribución normal.

Por otro lado, las variables hsizesq, hsize, hsrank, hsperc tiene diferencias considerables entre la media normal y media trim y winsorizada En el caso de la media normal, su cálculo es sensible a valores extremos. Por el contrario, los valores de la media trim y winsorizada, son muy similares, ya que su implementación es robusta y puede procesar los valores extremos.

Respecto a la variable tothrs los valores de la media normal y la mean_win y media_trim son similares, lo cual tiene sentido, dado que sus valores tienden a una distribución uniforme.

Para complementar el análisis anterior, se grafica los histogramas de las 8 variables continuas.

Se observa un aspecto interesante en la variable tothrs, su distribución no es precisamente una distribución uniforme sino una distribución multimodal.

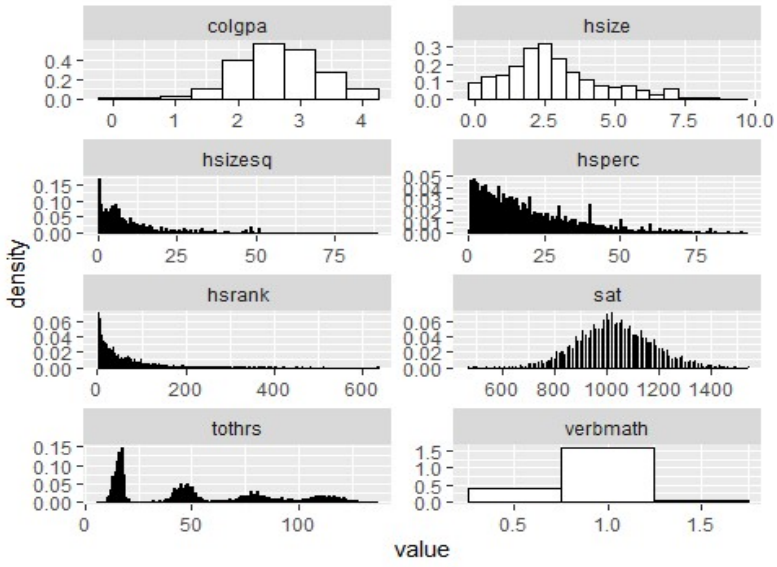

Figura 2. Histogramas variables continuas [fuente propia]

Continuando con el análisis se plantea un estudio multivariable, para ello se realiza un gráfico de dispersión.

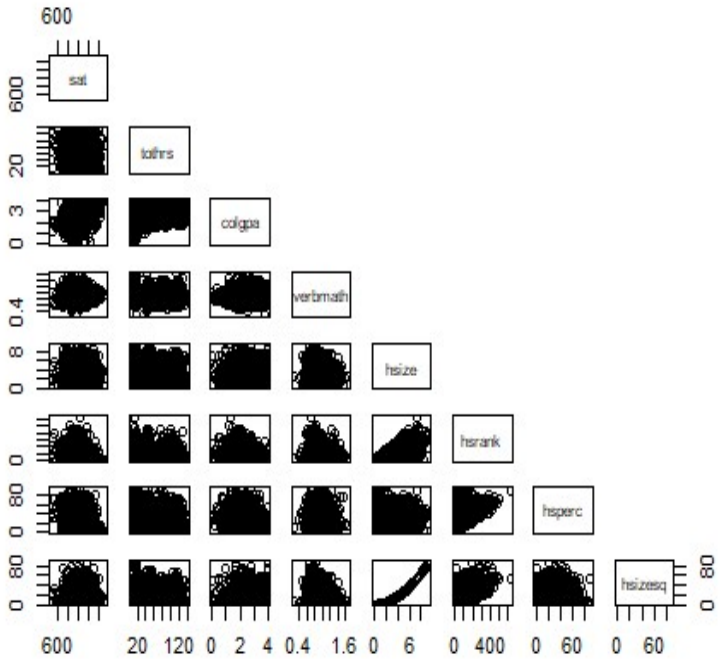

Figura 3. Grafico dispersión variables continuas [fuente propia]

En la figura se observa que las variables hsize y hsizesq, presentan un comportamiento exponencial. Esto sucede porque hsizesq, es el valor cuadrado de hsize. En menor medida se observa un comportamiento lineal entre las variables hsrank y hsperc con hsize. Para las demás variables, no se aprecia un comportamiento definido.

3.1.2. Test de normalidad

Se realiza test de normalidad, basado en la inspección de histogramas de las variables y gráficos Q-Q (ver figura 4), de los que se concluye lo siguiente:

- sat: sigue una distribución normal.
- tothrs: no sigue una distribución normal, presenta múltiples modas.
- colgpa: sigue una distribución normal.
- verbmath sigue una distribución normal.
- hize: no normal, tiene sesgo a la derecha.
- hsrank: no normal, tiene sesgo a la derecha.
- hsperc: no normal, tiene sesgo a la derecha.
- hisisq: no normal, tiene sesgo a la derecha.

Tambien se aplican los test de Kolmogorov-Smirnov y Shapiro-Wilk, para evaluar la normalidad de las variables, desde un enfoque cuantitativo.

Tabla 6. Test estadísticos para evaluar normalidad de variables [fuente propia]

Test	sat	tothrs	colgpa	verbmath	hsize	hsrank	hsperc	hsizesq
Kolmogorov-Smirnov	3.93E-04	0.00E+00	6.40E-02	8.14E-08	0.00E+00	0.00E+00	0.00E+00	0.00E+00
Shapiro-Wilk	4.42E-05	1.79E-46	1.85E-15	6.07E-25	1.47E-34	2.28E-63	1.56E-48	8.43E-61

Los dos test muestran que todos los p-value son menores que el nivel de significancia (0.05) y por ende se concluye que ninguna de las variables tiene una distribución normal.

No obstante, y tras considerar tanto el análisis gráfico y el teorema de límite central (el tamaño de la muestra para cada variable es de 4137 observaciones), se puede afirmar que al menos las variables sat, verbmath y colgpa, tienen una distribución normal.

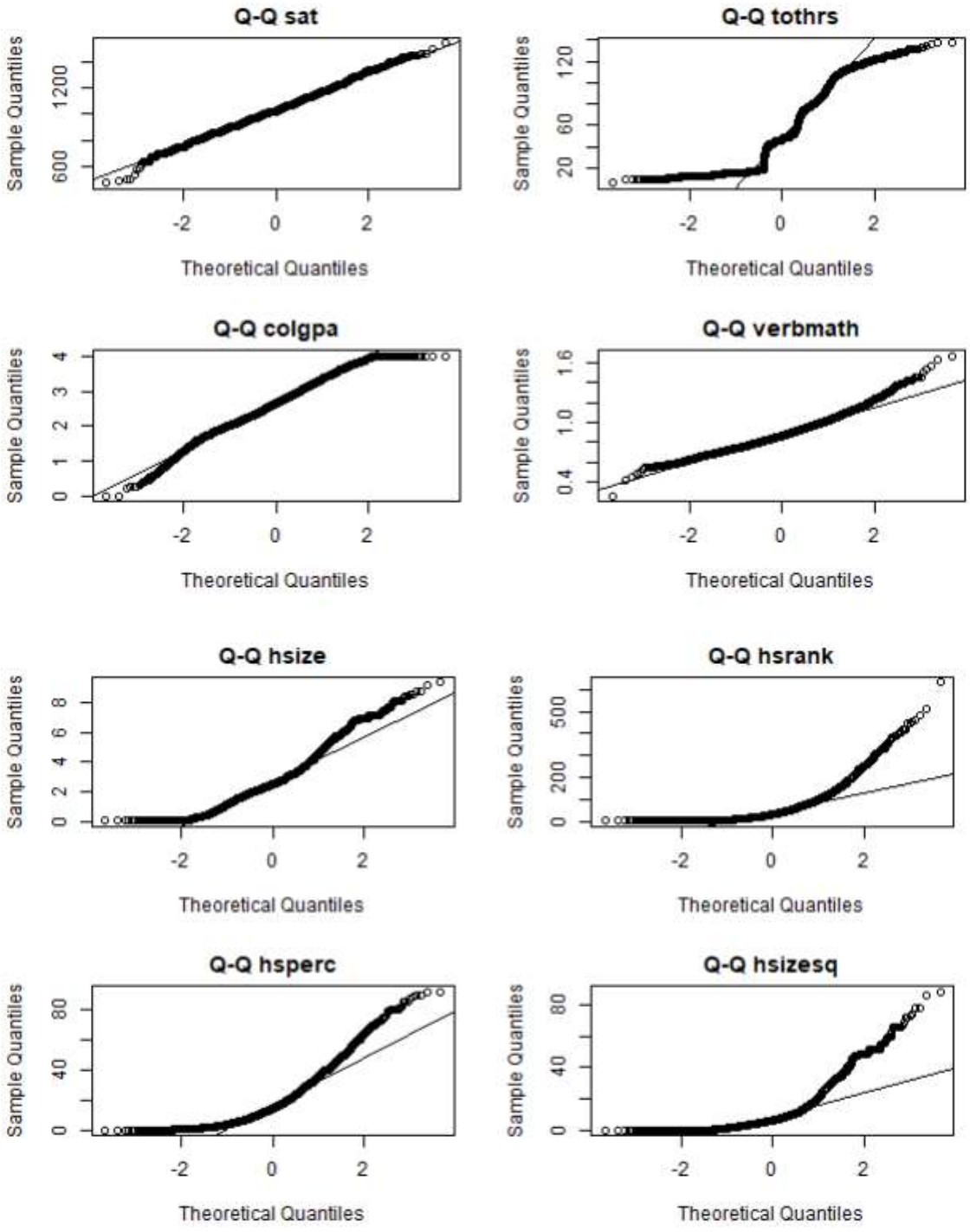

Figura 4. Gráficos Q-Q variables numéricas [fuente propia]

3.1.3. Análisis de correlación

Dado que la mayoría de las variables no siguen una distribución normal, se aplicará la técnica de correlación de spearman, en la cual no asume ninguna suposición sobre la distribución de los datos.

	sat	*tothrs*	*colgpa*	*verbmath*	*hsize*	*hsrank*	*hsperc*	*hsizesq*
sat								
tothrs	0.030							
colgpa	0.395***	0.144***						
verbmath	-0.004	0.004	0.016					
hsize	0.081***	-0.046**	-0.004	-0.050**				
hsrank	-0.234***	-0.119***	-0.393***	-0.031*	0.593***			
hsperc	-0.325***	-0.127***	-0.483***	-0.001	-0.024	0.713***		
hsizesq	0.081***	-0.046**	-0.004	-0.050**	1.000***	0.593***	-0.024	

Computed correlation used spearman-method with pairwise-deletion

Figura 5. Correlación de spearman y p-value de variables continuas numéricas [fuente propia]

Los valores en negrita sugieren que el p-value es significativo y en caso contrario que no es significativo. De acuerdo a lo anterior, se puede afirmar por ejemplo que la correlación usando el método de spearman para las variables colgpa y sat es de 0.395 con un nivel de confianza del 95%. Por el contrario, no se puede concluir que exista correlación entre verbmath y colgpa, dado que el p-value es mayor que 0.05.

A continuación, se crea gráfico de correlación para visualizar de forma más clara los resultados obtenidos.

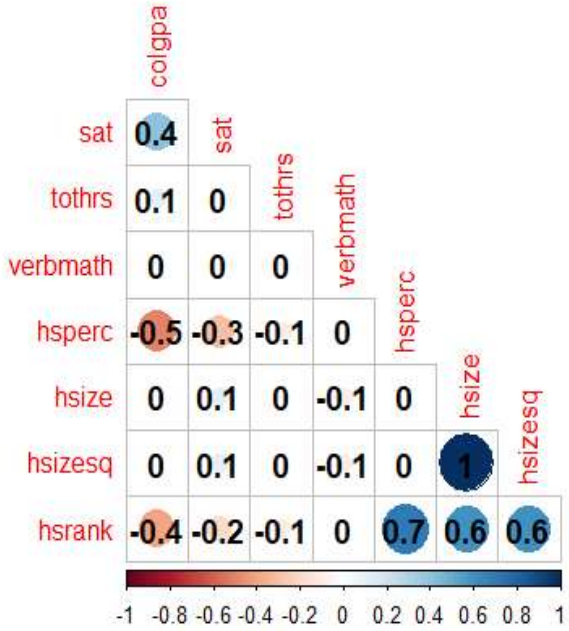

Figura 6. Correlación de spearman variable continuas numéricas [fuente propia]

Del gráfico y de los resultados de la sección anterior, se concluye que hay correlación positiva entre hsrank y hsperc, hsize. y negativa entre hsrank, hsperc y colgpa y sat, con un nivel de confianza del 95%.

3.1.4. Análisis de linealidad entre variables colgpa y sat en función de las variables continuas

Dado que el interés es buscar relaciones que expliquen los valores obtenidos para colgpa y sat, a continuación, se hace un análisis de dichas variables en función de las variables continuas restantes.

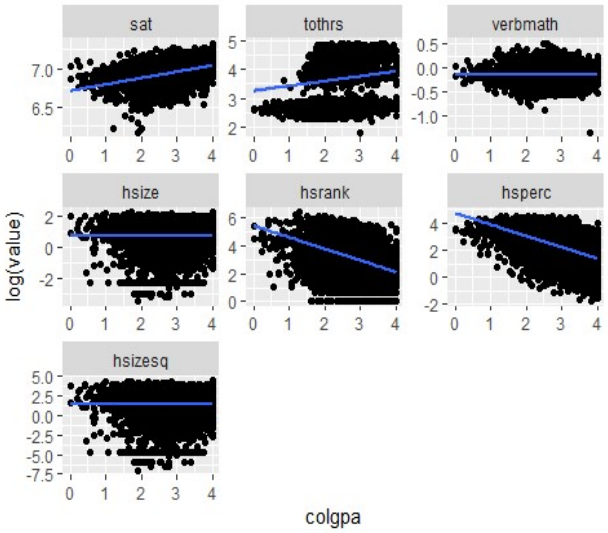

Figura 7. Relación lineal entre variables colgpa y variables continuas [fuente propia]

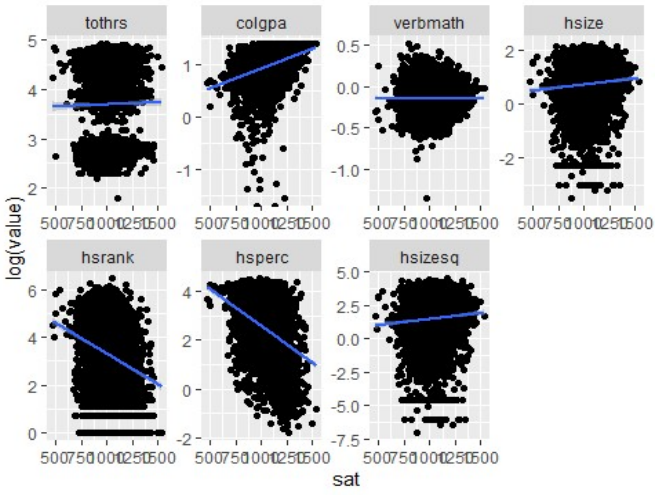

Figura 8. Relación lineal entre variables sat y variables continuas [fuente propia]

De los gráficos anteriores, se aprecia una relación lineal con pendiente positiva entre sat y colgpa, también una relación lineal con pendiente negativa entre hsperc, hsrank, respecto a sat y colgpa. En las variables tothrs, verbmath y hsize, se observa una relación lineal positiva mínima respecto sat y colgpa

3.2. Análisis de variables categóricas

En esta sección se realiza el análisis de las variables categóricas.

3.2.1. Análisis Descriptivo

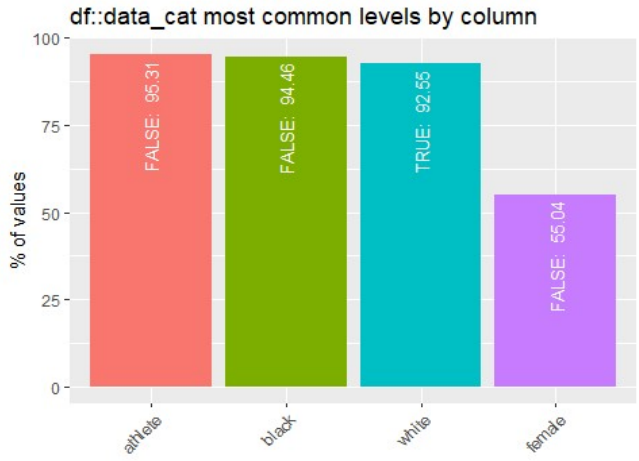

Figura 9. Proporción clases variables categóricas [fuente propia]

En la variable athlete se observa que el 95% de los datos están categorizados como falso, lo que indica un porcentaje alto de estudiantes que no son atletas. Respecto a la variable black se observa que un 94% de los datos están categorizados como falso, lo que indica que hay alto porcentaje de estudiantes de raza blanca. Respecto a la variable White, se observa que un 93% de los datos están categorizados como verdadero, lo que indica que hay un bajo porcentaje de estudiantes de raza negra. Finalmente, para la variable female, se aprecia un 55% de los datos están categorizados como falso, lo que indica que el 55% de los estudiantes son hombres.

4. Problema de estudio

En este apartado se proponen dos problemas que se estudiarán desde dos enfoques, el primero desde la inferencia estadística y el segundo desde la perspectiva de modelos predictivos supervisados.

4.1. Basado en estadística inferencial

En primera instancia se plantea el problema de determinar si el hecho de que un estudiante sea atleta influye en el promedio de la nota final.

El primer paso es seleccionar las variables de interés, para ello se selecciona la variable colgpa y se selecciona si el estudiante es o no atleta

```
colgpa_athlete <- data$colgpa[data$athlete == TRUE]
colgpa_no_athlete <- data$colgpa[data$athlete == FALSE]
```

Se grafica el histograma sobre el que se superpone el valor medio y también un gráfico Q-Q para evaluar visualmente si las muestras de las dos variables de interés siguen una distribución normal.

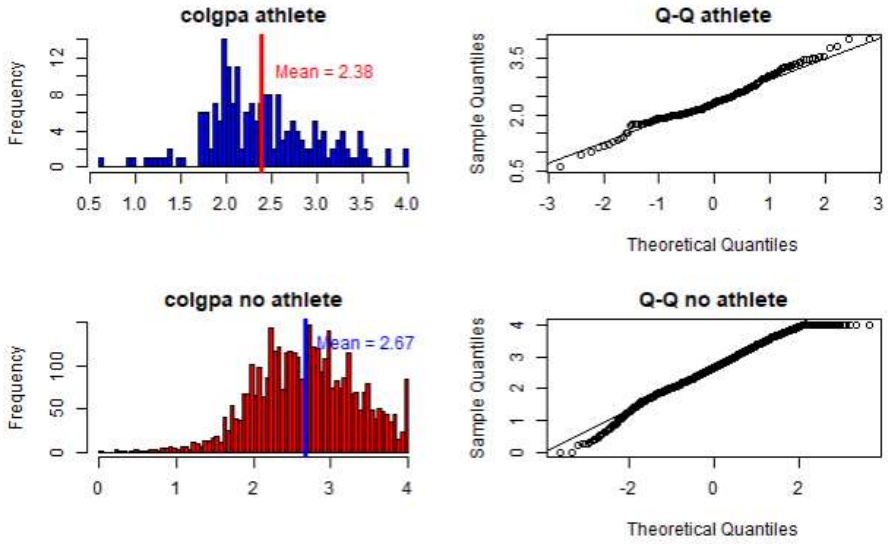

Figura 10. Histograma y grafico Q-Q muestras de análisis [fuente propia]

A partir de los gráficos Q-Q se aprecia que las dos muestras siguen una distribución normal Por otro lado se aprecia que el histograma también tiene la forma de una distribución normal. Otro aspecto interesante es que el valor promedio para la muestra colgpa no athlete, es mayor respecto a colgpa no athlete. Preliminarmente a partir de este enfoque descriptivo se puede inferir que el valor promedio de la nota en los estudiantes difiere si es o no atleta.

4.1.1. Análisis de normalidad de las muestras

Se realizan dos test estadísticos para evaluar la normalidad en las muestras (ver tabla 7). Para las dos muestras se obseva que los dos test son contradictorios. Por un lado con el test shapiro-Wilk se concluye que las muestra no siguen una distribución normal y por otro lado, con el test de kolmogorov-Smirnov se concluye que sí.

Tabla 7. Test normalidad para las muestras de estudio [fuente propia]

Test	colgpa_athlete	colgpa_no_athlete
Shapiro-Wilk	0.01153	0.00000
Kolmogorov-Smirnov	0.23340	0.05364

Aunque los test son contradictorios, de acuerdo al análisis gráfico y considerando el teorema de límite central, se puede considerar que las dos muestras tienen una distribución normal.

4.1.2. Análisis de la varianza entre las muestras

Tras realizar el test de varianza se aprecia que el p-value es mayor que 0.05, con lo cual se infiere que las dos muestras tienen varianzas iguales.

```
var_test <- var.test(colgpa_athlete, colgpa_no_athlete)
var_test$p.value = 0.07653321
```

4.1.3. Pregunta de Investigación

Para abordar el problema de forma específica, se plantea la siguiente pregunta:
¿Hay diferencias en el promedio de nota final de los estudiantes si son o no atletas?

4.1.4. Hipótesis nula y alternativa

Dado que es un problema de inferencia estadística, se deben plantear la hipótesis nula y alternativa. La hipótesis nula plantea que no hay diferencia en el promedio de nota final por el hecho que un estudiante sea atleta y la hipótesis alternativa plantea que si hay diferencia en el promedio de nota final.

$$H_0: \ u_1 = u_2$$
$$H_1: \ u_1 \neq u_2$$

4.1.5. Justificación del test a aplicar

El test a aplicar es el **contraste de dos muestras independientes sobre la media con varianzas desconocidas pero iguales.**

.
A continuación, se plantean la justificación de porque se considera adecuado dicho test para resolver el problema.

- El tamaño de las muestras es diferente y no tienen datos faltantes/nulos, por ende se considera que las dos muestras son independientes.

- Se asume que las dos muestras siguen una distribución normal, de acuerdo al análisis gráfico y de acuerdo al teorema de limite central (colgpa_athlete = 194 y colgpa_no_athlete = 3943).

- Las varianzas son iguales, lo cual se comprobó con el test de varianza

- Se desconoce la varianza poblacional.

4.1.6. Cálculo del test

El test de hipótesis se realiza mediante la función t.test.

```
t.test(colgpa_athlete, colgpa_no_athlete, var.equal=TRUE)
```

```
 Two Sample t-test
data:  colgpa_athlete and colgpa_no_athlete
t = -5.8983, df = 4135, p-value = 3.966e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3791097 -0.1899574
sample estimates:
mean of x mean of y
 2.381495  2.666028
```

4.1.7. Interpretación

De acuerdo al test anterior, se concluye que si hay diferencias en las notas de los estudiantes si estos son o no atletas, con un nivel de confianza del 95%. Lo anterior se infiere a partir del p-value, cuyo valor es menor que 0.05. De igual manera, el tobs está fuera de la región de aceptación, con lo cual se rechaza la hipótesis nula.

4.2. Basado en modelos supervisados

En este este apartado se abordará el problema: **¿Se puede predecir la nota promedio final de los estudiantes?**, al cual se le dará respuesta desde la perspectiva de los modelos supervisados, en particular la regresión lineal y árboles de decisión con técnica de ensamble.

4.2.1. Modelo de regresión lineal

En primer lugar, se seleccionan las variables de interés y se aplican transformaciones para mejorar algunas propiedades estadísticas, con lo cual se brinde una mayor capacidad predictiva/generalización al modelo.
En este sentido, se aplican transformaciones a las variables hsperc, hsrank, hsizesq y tothrs, mediante el test BoxCox. En el siguiente gráfico se corrobora que la distribución de las transformaciones, tuvo un cambio significativo evidenciando que las gráficas tienden hacia una distribución normal.

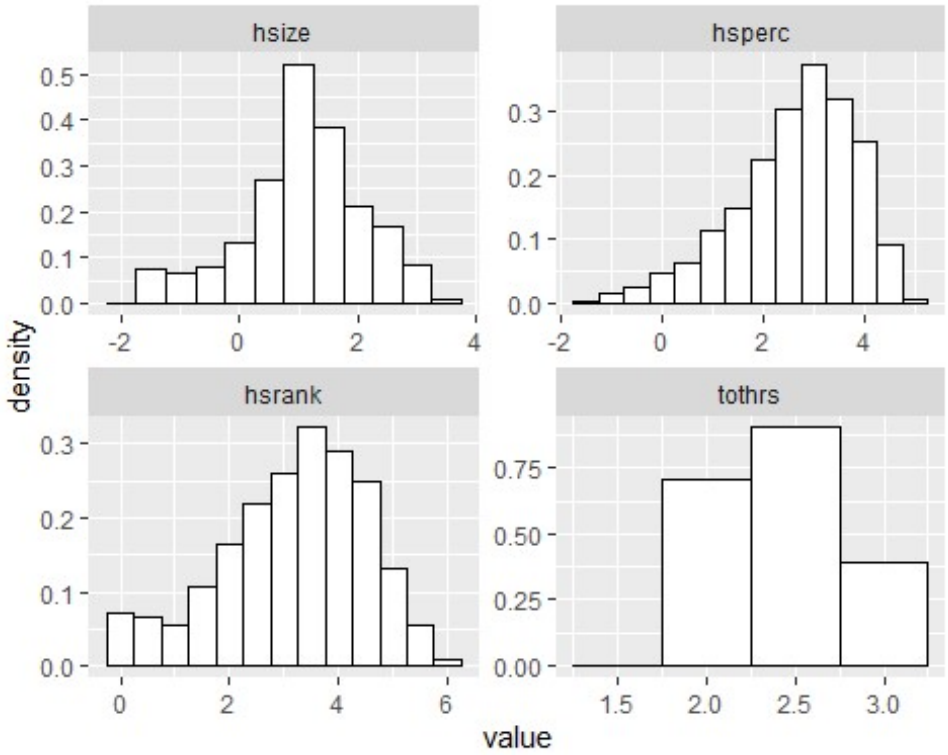

Figura 11. Distribuciones luego de aplicar transfomración [fuente propia]

También se prescindirá de la variable hisizesq, por su alta correlación con la variable hisize, con lo cual se elimina la información redundante y se disminuye el efecto de la multicolinealidad

Se divide el conjunto de datos en la proporción 70% para entrenamiento y 30% para test.

```
set.seed(123)
split <- createDataPartition(data$colgpa, p= .7, list = FALSE, times = 1)
train <- data[split,]
test <- data[-split,]
```

Con la función lm se crea modelo de regresión lineal multivariado.

```
Model1 <- lm(colgpa ~., data=train)
summary(Model1)

 Call:
 lm(formula = colgpa ~ ., data = train)

 Residuals:
     Min       1Q   Median       3Q      Max
-2.17941 -0.35331  0.01654  0.36993  1.90546

 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  1.613e+00  1.549e-01  10.417  < 2e-16 ***
sat          1.226e-03  8.352e-05  14.681  < 2e-16 ***
tothrs       1.744e-01  3.002e-02   5.811 6.88e-09 ***
athleteTRUE  1.314e-01  5.013e-02   2.621  0.00882 **
verbmath    -1.377e-01  6.868e-02  -2.005  0.04502 *
hsize       -1.608e-02  3.617e-02  -0.444  0.65677
hsrank      -2.087e-02  4.194e-02  -0.498  0.61882
hsperc      -1.972e-01  3.687e-02  -5.347 9.63e-08 ***
femaleTRUE   1.494e-01  2.123e-02   7.034 2.49e-12 ***
whiteTRUE    1.865e-02  7.552e-02   0.247  0.80495
blackTRUE   -3.423e-01  8.674e-02  -3.946 8.13e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5372 on 2887 degrees of freedom
Multiple R-squared:  0.3346, Adjusted R-squared:  0.3323
F-statistic: 145.2 on 10 and 2887 DF,  p-value: < 2.2e-16
```

Los resultados muestran que el modelo tiene una baja capacidad predicativa, dado que el valor de Multiple R-squared (R2) es de 0.335. También se aprecia que las variables verbmath, hsize, hsrank y white, no aportan información relevante al modelo, ya que los valores del p-value son mayores que 0.05. En contraste, variables con sat, tothrs, athlete, hsperc y female, aportan información relevante, dado que los p-value son menores que 0.05.

Para evaluar el desempeño del modelo en conjunto de test, se usa la función predict

```
colgpa_pred1 <- predict(Model1, newdata=test, type='response')
R2(colgpa_pred1, test$colgpa, form = "traditional")
0.3664266
```

Como se aprecia el valor de R2 mejora levemente, sin embargo, sigue teniendo una baja capacidad predictiva.

Se genera un nuevo modelo, pero se entrenará con las variables que aportan más información, es decir, sat, tothrs, athlete, hsperc y female.

```
Model2 <- lm(colgpa ~., data=train)
summary(Model2)

 Call:
 lm(formula = colgpa ~ ., data = train)

 Residuals:
     Min       1Q   Median       3Q      Max
-2.18675 -0.35735  0.02004  0.37053  1.91287

 Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.477e+00  1.279e-01  11.545  < 2e-16 ***
sat          1.211e-03  8.286e-05  14.614  < 2e-16 ***
tothrs       1.799e-01  3.003e-02   5.991 2.34e-09 ***
athleteTRUE  1.300e-01  4.983e-02   2.609  0.00914 **
```

```
hsperc      -2.141e-01  9.457e-03 -22.639  < 2e-16 ***
femaleTRUE   1.423e-01  2.086e-02   6.823 1.08e-11 ***
blackTRUE   -3.546e-01  4.675e-02  -7.585 4.44e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5383 on 2891 degrees of freedom
Multiple R-squared:  0.3311, Adjusted R-squared:  0.3298
F-statistic: 238.5 on 6 and 2891 DF,  p-value: < 2.2e-16
```

Se aprecia una leve mejora en el valor de Multiple R-squared (R2), sin embargo, el modelo tiene una capacidad predictiva muy baja.

A continuación, se realiza validación del modelo en el conjunto de entrenamiento, donde se observa una leve mejora en el valor de R2. Tal como en el caso anterior, el modelo no tiene una buena capacidad predictiva.

```
colgpa_pred2 <- predict(Model2, newdata=test, type='response')
R2(colgpa_pred2, test$colgpa, form = "traditional")
0.3630421
```

4.2.1. Modelo de árbol de decisión y ensamble

En este apartado se aborda un enfoque más robusto para intentar dar respuesta a la pregunta. Para ello se usará como modelo base los arboles de decisión y se usará una técnica de ensamble denominada baggging. En dicha técnica se generan múltiples modelos de árboles de decisión, a partir de la selección aleatoria con reemplazó de las features o variables independientes. Posteriormente se realiza un proceso de votación asignando el mismo peso a cada modelo generado y finalmente se toma el valor promedio de todos los modelos generados.
Con el propósito de entrenar el modelo de forma robusta y disminuir la sobreoptimizacíon, se usa la técnica de validación cruzada; para este caso se selecciona el método k-fold, con una partición del conjunto de entrenamiento en 10 folds.

```
set.seed(123)

ctrl <- trainControl(method = "cv",  number = 10)

Model3 <- train(
  colgpa ~ .,
  data = train,
  method = "treebag",
  trControl = ctrl,
  importance = TRUE
  )

Model3
 Bagged CART
```

```
2898 samples
  10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2608, 2608, 2608, 2607,
2608, 2609, ...
Resampling results:

  RMSE       Rsquared  MAE
  0.5493791  0.304804  0.4366664
```

Los resultados muestran que el valor de R2 es de 0.304804, que es muy bajo y por ende se concluye que el modelo tiene baja capacidad predictiva.

Otro aspecto interesante que se puede extraer del modelo, son las variables que aportan mayor información y capacidad predictiva. Tal como se aprecia en la figura, las tres variables que aportan mayor información son **sat**, **hsperc** y **hsrank**, que es coherente con los valores obtenidos con el modelo de regresión lineal.

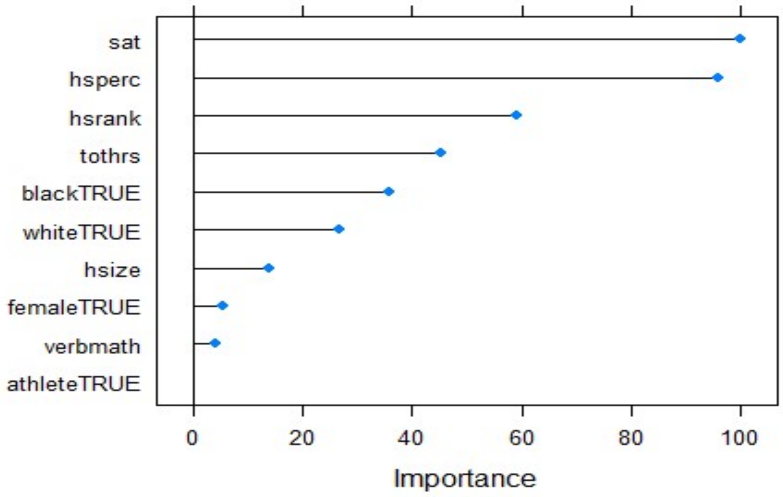

Figura 12. Importancia de las variables en el modelo [fuente propia]

Finalmente se evalúa la capacidad predictiva/generalización del modelo en el conjunto de test. Tal como aprecia hay una leve mejora en el valor de R2, sin embargo, el modelo tiene una baja capacidad predictiva.

```
colgpa_pred3 <- predict(Model3, test)
postResample(colgpa_pred3, test$colgpa)
Rsquared: 0.3312981
```

4.2.3. Interpretación:

De acuerdo al resultado obtenido, con los modelos de regresión lineal y árbol de decisión con la técnica bagging, no se puede dar respuesta al problema de estudio, ya que la precisión de los modelos, tanto el conjunto de entrenamiento y de test es muy baja, en promedio el valor de R^2 para los tres modelos es de 0.30, lo cual no permite dar una respuesta concreta al problema de estudio.

5. Código

Para acceder al código se debe ingresar a:

https://github.com/jhontd03/gpaanalysis.git

6. Contribuciones

Contribuciones	Firma
Investigación previa	Jhon Jairo Realpe
Redacción de las respuestas	Jhon Jairo Realpe
Desarrollo del código	Jhon Jairo Realpe
Participación en el vídeo	Jhon Jairo Realpe

7. Enlace Video

El video se puede visualizar en:

https://drive.google.com/file/d/1HJv39C1xLwtjQIYD3KTcx8ANpm9LTIH_/view?usp=share_link

Bibliografía

College Board Sat Program. (2022). SAT Understanding Scores. Miami. Retrieved from https://satsuite.collegeboard.org/media/pdf/understanding-sat-scores.pdf

Econpapers. (2000). gpa2. Retrieved January 11, 2023, from https://econpapers.repec.org/paper/bocbocins/gpa2.htm

Forstall, M. (2019). What Are Verbal SAT Scores? Retrieved January 12, 2023, from https://www.theclassroom.com/verbal-sat-scores-8525646.html

Muniz, H. (2021). SAT Score Range: 3 Steps to Understanding Your Score. Retrieved January 12, 2023, from https://blog.prepscholar.com/sat-score-range

PrepScholar. (2018). Coe College SAT Scores and GPA. Retrieved January 12, 2022, from https://www.prepscholar.com/sat/s/colleges/Coe-College-SAT-scores-GPA

Zhang, F. (2018). SAT Score to GPA Conversion Table. Retrieved January 12, 2022, from https://blog.prepscholar.com/sat-gpa-conversion-table