Trabajo estadística

Descripción estadística de una variable

Prueba de hipótesis y regresión lineal

Realizado por:

Manuela Herrera López

Samuel Palacios Bernate

Mónica Liliana Arteaga Sierra

Estadística general

Universidad EAFIT

Medellín

2020

Ingeniería de sistemas

Índice

Introducción:

La prueba saber 11 es un examen que se le aplica a todos los estudiantes de onceavo grado en Colombia; en la que se miden diferentes competencias y el nivel de aprendizaje adquirido durante los años de estudio. En la prueba, se evalúan las áreas de matemáticas, lectura crítica, ciencias sociales y ciudadanas, ciencias naturales e inglés, y por cada área se saca un puntaje por estudiante, lo que le dará una posición con respecto a los otros que desarrollaron la prueba. A partir de los resultados obtenidos no sólo se puede clasificar a los estudiantes, sino, también a sus instituciones educativas, además, permite visualizar cómo es la calidad de la educación en el Valle de Aburrá.

En el presente documento se analizará una base de datos que contiene los resultados por institución educativa de las pruebas saber 11, que van desde el período 2014-2 hasta el 2018-2. De él se extraerán diferentes datos, como los correspondientes a las comunas en las que están los establecimientos educativos y se compararán con respecto a los niveles de inglés obtenidos, se analizarán los puntajes globales por colegio y su relación con respecto al año de la prueba y así, ver cómo cambian. A su vez, se calcularán algunos índices de dispersión para saber cuán homogéneo es el aprendizaje de los jóvenes en el Valle de Aburrá y poder clasificarlos según algunos criterios posteriormente explicados.

Descripción de las variables:

Nombre	Tipo	Descripción
año_semestre	Categórica (ordinal)	Año y semestre de los resultados
establecimiento	Categórica (nominal)	Nombre del establecimiento educativo
matriculados	Numéricas (discreta)	Cantidad de matriculados
registros	Numéricas (discreta)	Cantidad de registrados
presentes	Numérica (discreta)	Número de presentes
puntaje_global	Numérica (discreta)	Puntaje global
puntaje_lectura	Numérica (discreta)	Puntaje de lectura
puntaje_matematicas	Numérica (discreta)	Puntaje de matemáticas
puntaje_sociales	Numérica (discreta)	Puntaje de sociales
puntaje_naturales	Numérica (discreta)	Puntaje de ciencias naturales
puntaje_ingles	Numérica (discreta)	Puntaje de inglés
comuna	Categórica (nominal)	Comuna o Corregimiento
prestacion_servicio	Categórica (nominal)	Prestación de servicio educativo

Análisis descriptivo de las variables en Python:

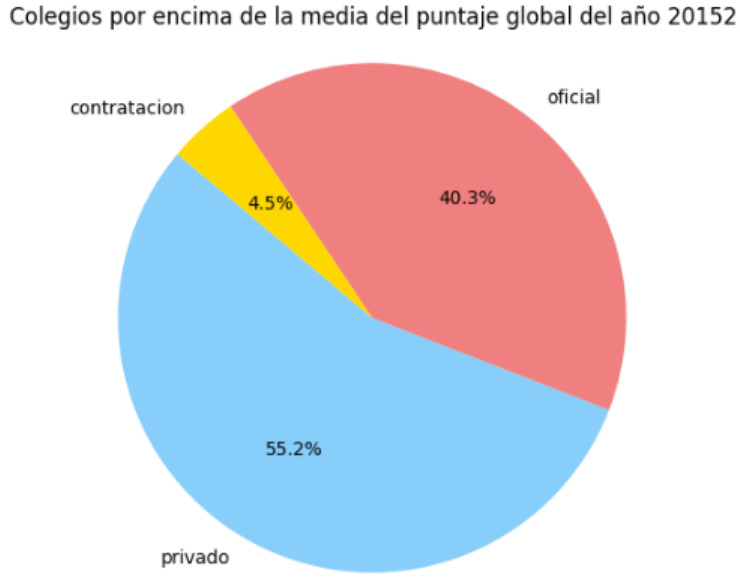
- **Colegios por encima de la media de puntaje global:**

  Revisando la gráfica de torta para los años desde 2014 hasta 2018 en calendario A, podemos notar que la proporción de institutos privados que obtienen resultados por encima de la media para el puntaje global es mayor que la de institutos públicos o de contratación, lo que nos indica que la calidad de educación es mejor en estos establecimientos con base en los resultados obtenidos.

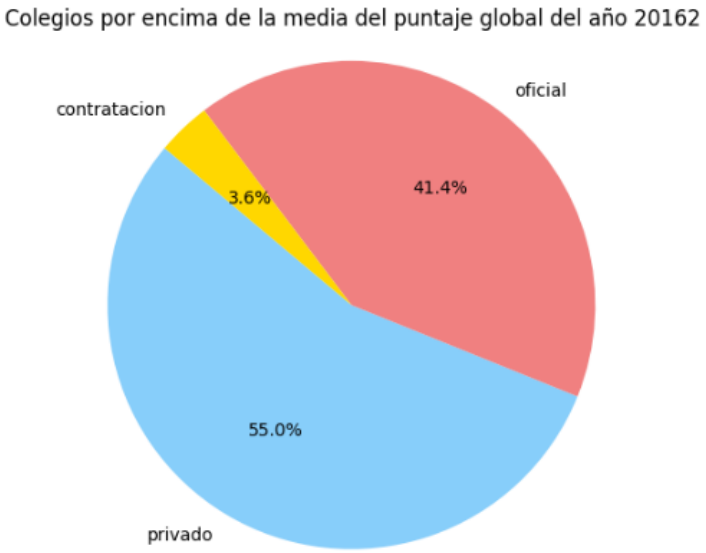
  A pesar de lo anterior, se puede evidenciar que la cantidad de colegios oficiales que tienen un puntaje global igual o mayor que la media desde el 2014-2 ha incrementado, lo que puede significar que la calidad de la educación ha mejorado en este tipo de institución educativa.

  Período 2014-2

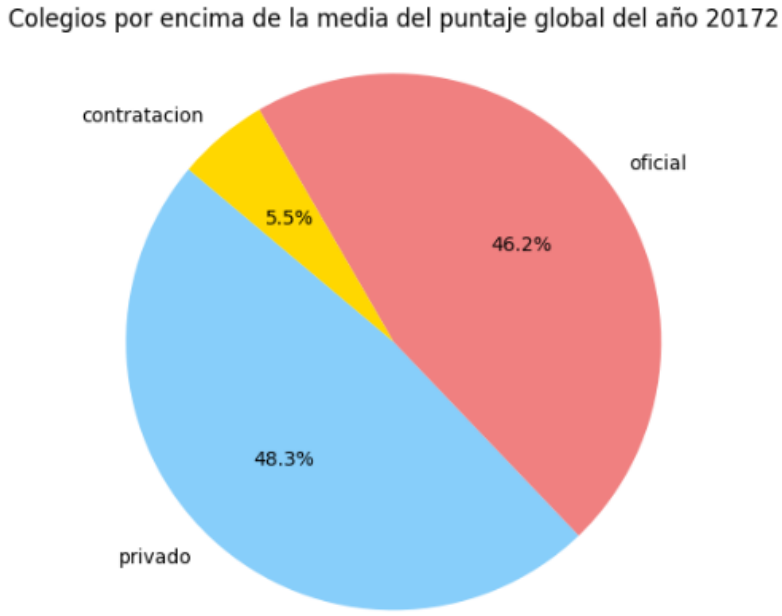
  Período 2015-2



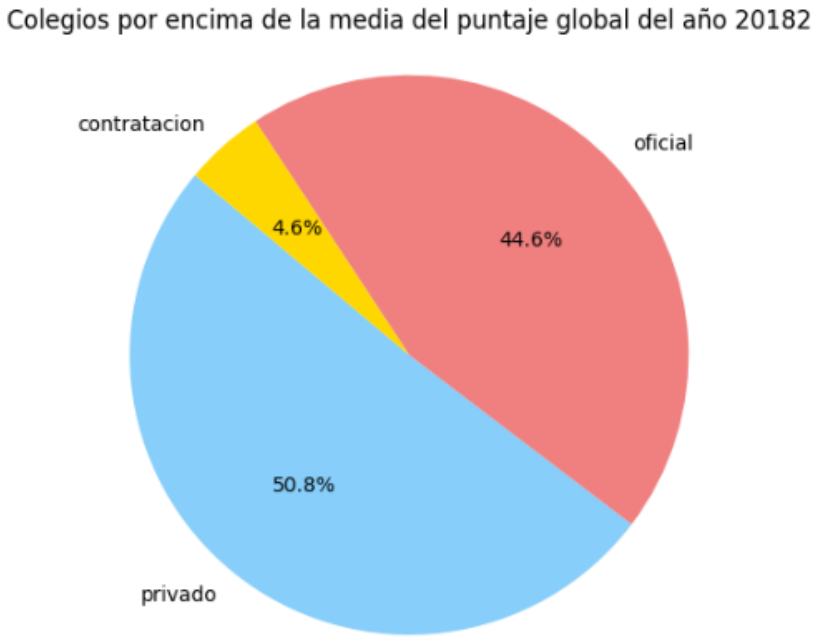
  Período 2016-2



Período 2017-2



Período 2018-2



- **Desviación estándar por años:**

  Con los datos de la variación estándar podemos ver cuál es el valor de la dispersión de la población con respecto al puntaje global de la prueba. Así, desde la media puede variar en más o menos el valor de la desviación estándar. O sea, que vemos que desde el 2014-2 hasta el 2018-2, el valor del puntaje global se aleja hacia arriba y hacia abajo en la proporción de 27,5 puntos de la media.

  Período 2014-2

  → General: 26.25
  → Instituciones privadas: 29.2
  → Instituciones públicas: 13.2

  Período 2015-2

  → General: 29.40
  → Instituciones privadas: 33.1
  → Instituciones públicas: 15.8

  Período 2016-2

  → General: 27.51
  → Instituciones privadas: 29.5
  → Instituciones públicas: 16.2

  Período 2017-2

  → General: 27.32
  → Instituciones privadas: 31.3
  → Instituciones públicas: 17.7

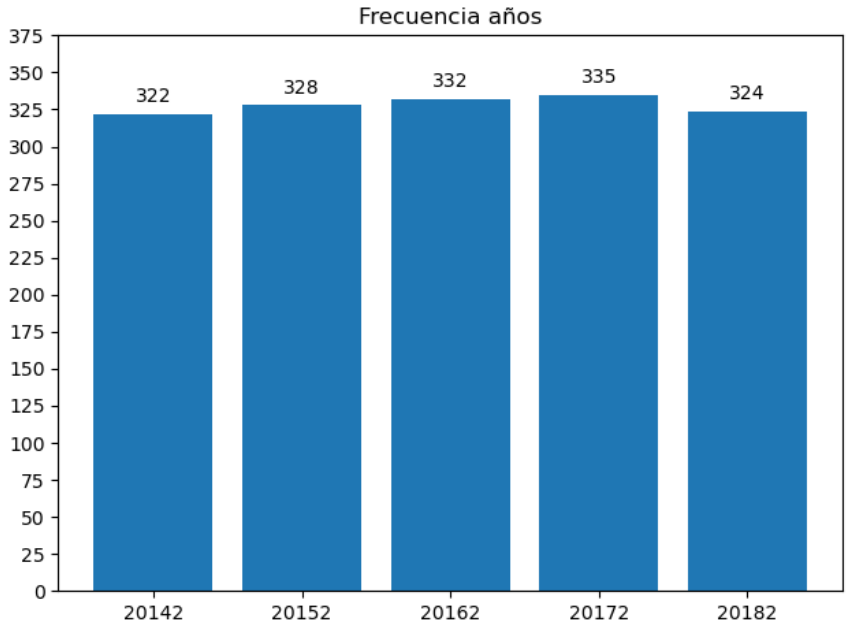
Período 2018-2

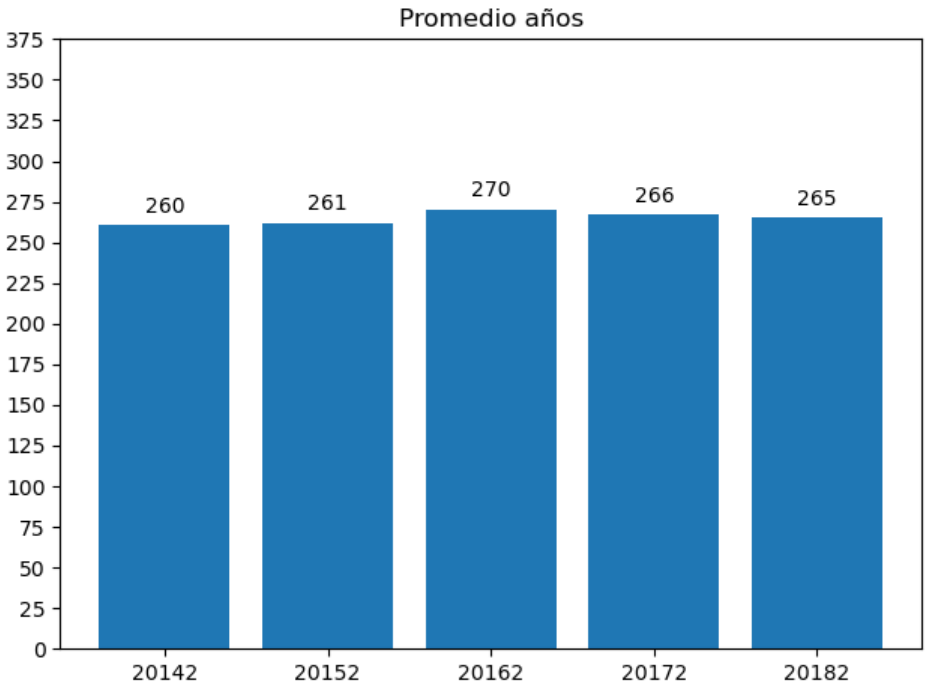
→ General: 27.05
→ Instituciones privadas: 29.2
→ Instituciones públicas: 17.7

Para profundizar en el análisis de la desviación estándar, vamos a ver cómo cambia entre los colegios públicos y privados del Valle de Aburrá. Gracias a estos valores podemos ver que, en los colegios privados, el puntaje global varía en mayor medida que en el de los colegios públicos, lo que implicaría que la educación es menos homogénea, pues hay un rango más amplio de la varianza de los resultados. O sea que las diferencias entre los desempeños de los estudiantes son mucho mayores.

- **Frecuencia puntaje por Año-Semestre:**

Al momento de comparar la tabla de frecuencia y promedio podemos apreciar que el puntaje global por año se mantiene en un rango estable de 260 a 270, lo cual indica que, aunque se aumentó el número de instituciones en las pruebas no se presentó una mejoría en relación a la calidad de educación en el Valle de Aburrá.





- **Teorema de Chebyshev**

Vamos a ver cuál es el porcentaje de los datos que está a una distancia de 2 desviaciones estándar con el teorema de Chebyshev, de lo que obtenemos que:
$1 - (1/2^2) = 0.75$, por lo que al menos el 75% de los datos están a 2 desviaciones estándar de la media.

→ Para el **2014-2**: del gráfico anterior vemos que la media es 260, y que la desviación estándar correspondiente es de 26.25, así, lo que sacamos con Chebysev nos dice que 26.25*2=52.5, lo que indica que el 75% de los colegios tienen un puntaje entre 207.8 y 312.5.

→ Para el **2015-2**: del gráfico anterior vemos que la media es 261, y que la desviación estándar correspondiente es de 29.4, así, lo que sacamos con Chebysev nos dice que 29.4 *2=58.8, lo que indica que el 75% de los colegios tienen un puntaje entre 202.2 y 319.8.

→ Para el **2016-2**: del gráfico anterior vemos que la media es 270, y que la desviación estándar correspondiente es de 27.51, así, lo que sacamos con Chebysev nos dice que 27.51*2=55, lo que indica que el 75% de los colegios tienen un puntaje entre 215 y 325.8.

→ Para el **2017-2**: del gráfico anterior vemos que la media es 266, y que la desviación estándar correspondiente es de 27.3, así, lo que sacamos con Chebysev nos dice que 27.3*2=54.6, lo que indica que el 75% de los colegios tienen un puntaje entre 211.4 y 320.6.

→ Para el **2018-2**: del gráfico anterior vemos que la media es 265, y que la desviación estándar correspondiente es de 27, así, lo que sacamos con Chebysev nos dice que 27*2=54, lo que indica que el 75% de los colegios tienen un puntaje entre 211 y 319.

- **Nivel de inglés por comunas:**

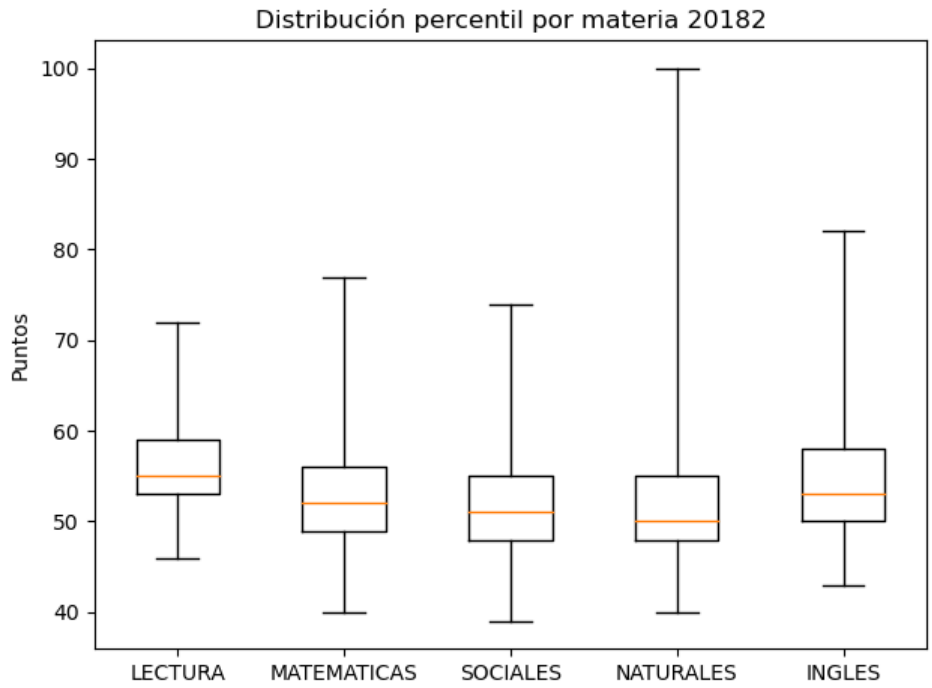
  En la siguiente tabla apreciamos una comparativa entre la media y la mediana de puntaje de inglés por comuna, esto con el propósito de saber si tenemos datos atípicos que puedan alterar la interpretación de los mismos.

→ Evidenciamos que las medidas no varían en un gran porcentaje, por lo que se puede afirmar que, por comunas, el nivel promedio de inglés de los colegios es aproximadamente el mismo, de tal forma que no hay grandes brechas de desigualdad o heterogeneidad en las poblaciones.
→ De la tabla podemos ver que los mejores puntajes en inglés están en los colegios en la zona del Poblado, y en segundo lugar de Laureles.
→ También se puede apreciar que los niveles más bajos en esta materia están en los colegios de los sectores del Popular y San Cristóbal.

	PUNTAJE_INGLES	
	Media	Mediana
POPULAR	48.318	48
SANTA CRUZ	49.345	48
MANRIQUE	49.019	49
ARANJUEZ	52.574	52
CASTILLA	52.881	52
DOCE DE OCTUBRE	50.165	50
ROBLEDO	53.554	52
VILLA HERMOSA	48.927	48
BUENOS AIRES	54.96	54
LA CANDELARIA	52.941	53
LAURELES ESTADIO	66.8	68
LA AMERICA	58.556	57
SAN JAVIER	49.712	48
POBLADO	70.126	69
GUAYABAL	55.24	55
BELEN	55.748	54
PALMITAS	51.6	52
SAN CRISTOBAL	48.569	48
ALTAVISTA	52.2	49
SAN ANTONIO DE PRADO	51.772	52
SANTA ELENA	54.6	55

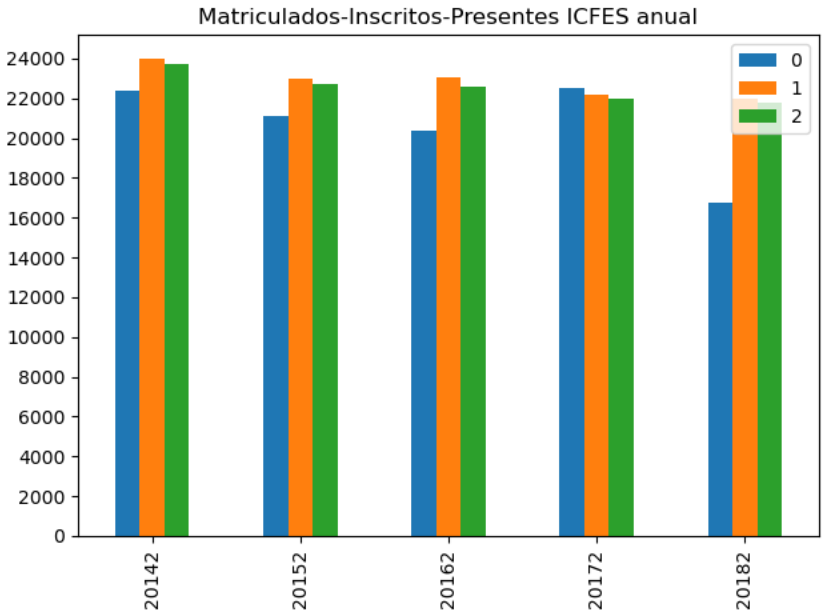
- **Diagrama de bigotes por materia en el período 2018-2**

→ Dado el diagrama de bigotes podemos inferir varias cosas, inicialmente se puede ver cómo la mediana varía dependiendo de la materia, lo cual indica que el aprendizaje no es homogéneo en todas las asignaturas. A su vez podemos ver que los cuartiles desde 25 a 75 en todas las asignaturas están entre 50 y 60 puntos, lo que implica que la mayoría de la población se encuentra está por debajo de los 60 puntos por materia.
→ Podemos ver que hay una varianza significativa en los extremos superiores por materia, el mejor desempeño está en ciencias naturales, seguido de inglés y matemáticas. Por otro lado, la distribución del extremo inferior, o sea el menor dato de ciencias, matemáticas y sociales se encuentra muy alineado.
→ Vemos que son asimétricas, y en la mayoría de los casos, los puntajes están por encima de la media, lo que es un buen indicador.

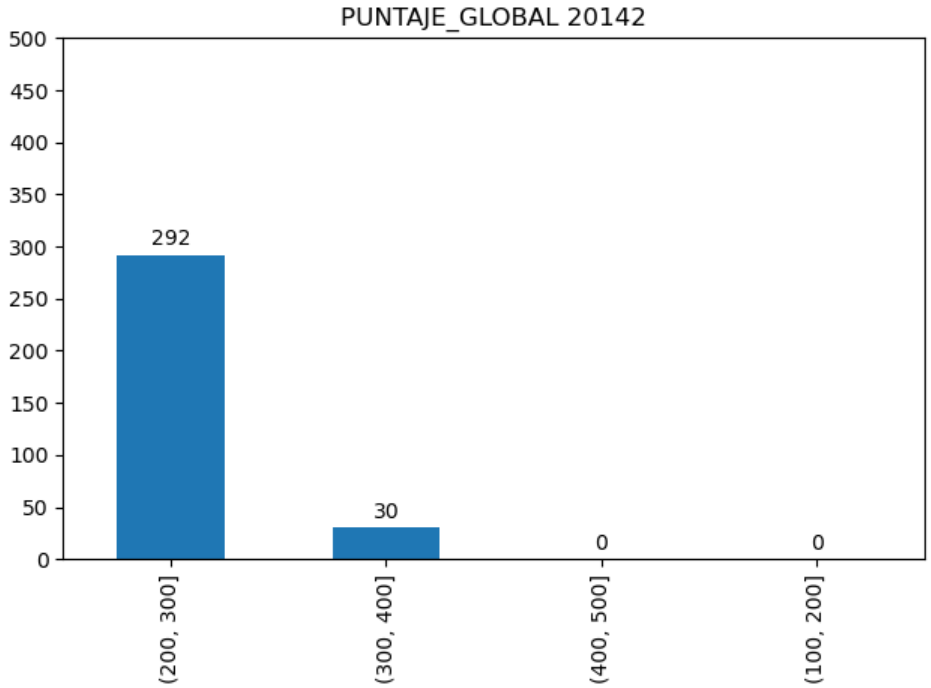


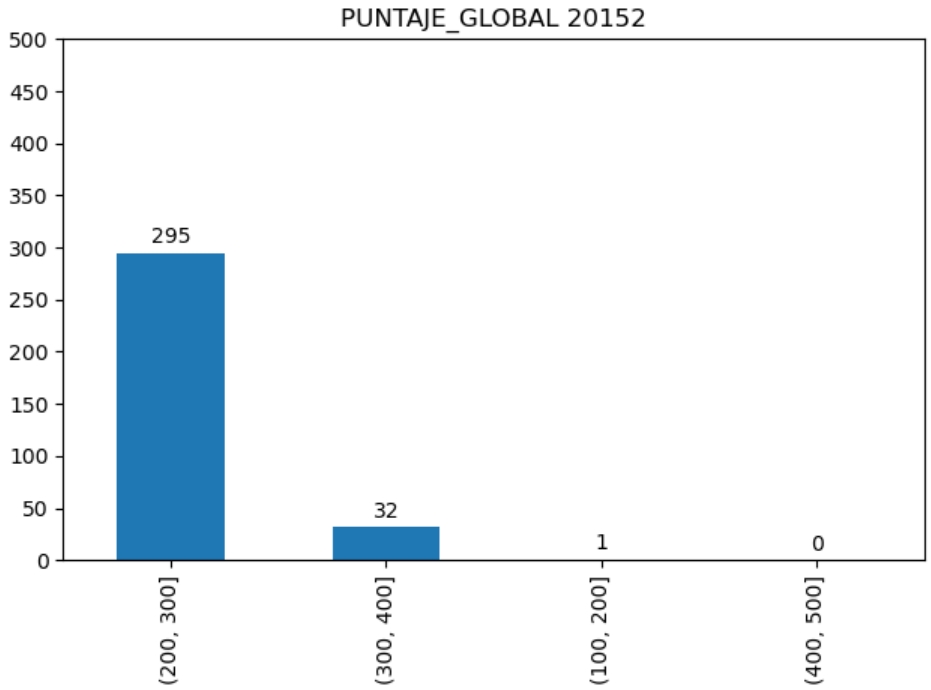
- **Estudiantes matriculados, inscritos y presentes**
  Matriculados (0): Estudiantes registrados en el Sistema Integrado de Matrícula (SIMAT) de acuerdo a la fecha de corte definida.
  Inscritos (1): Estudiantes enlistados para la presentación del examen.
  Presentes (2): Inscritos que asistieron a las dos sesiones del examen.

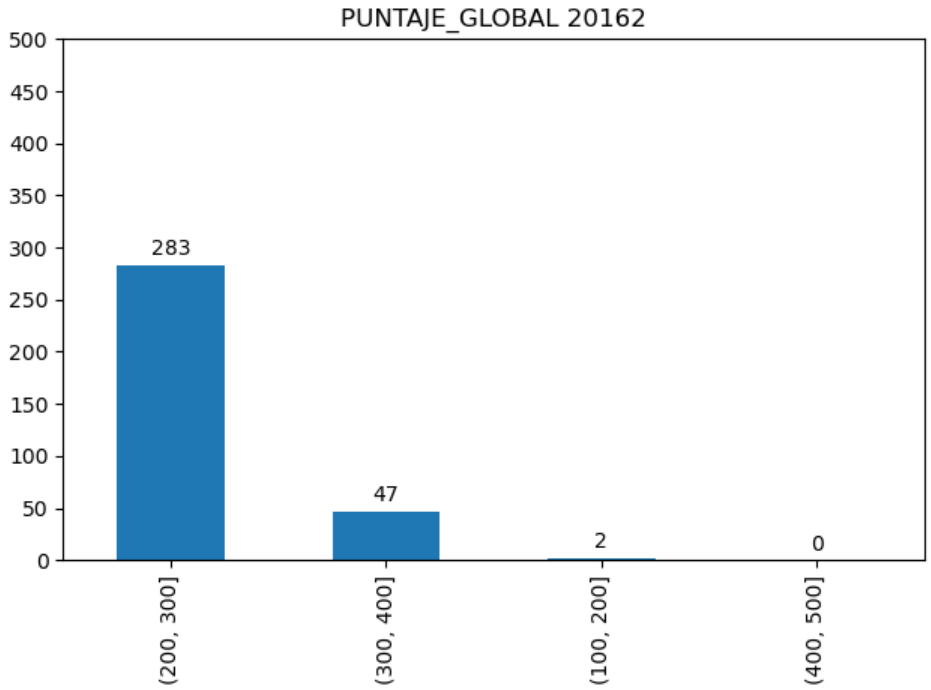
  Cuando agrupamos la suma de estos valores por cada año podemos evidenciar varias cosas:
  - → El índice de estudiantes matriculados habitualmente es menor al de inscritos, esto nos muestra que una buena parte de la población no se registra a tiempo para presentar sus pruebas ICFES.
  - → El índice de estudiantes que se presentan a las pruebas no es igual al de estudiantes registrados por lo que inferimos un pequeño porcentaje de incumplimiento independientemente del motivo.
  - → A diferencia del periodo 2017-2, en el que el número de matriculados fue mayor al de inscritos, lo que indica que los estudiantes se registraron a tiempo, en el período 2018-2, apenas un año después, la cantidad de matriculados estuvo muy por debajo de lo habitual en contraste con el número de registrados ese mismo periodo, lo que implica que un alto porcentaje de estudiantes se inscribieron después de la fecha de corte establecida por el ICFES.

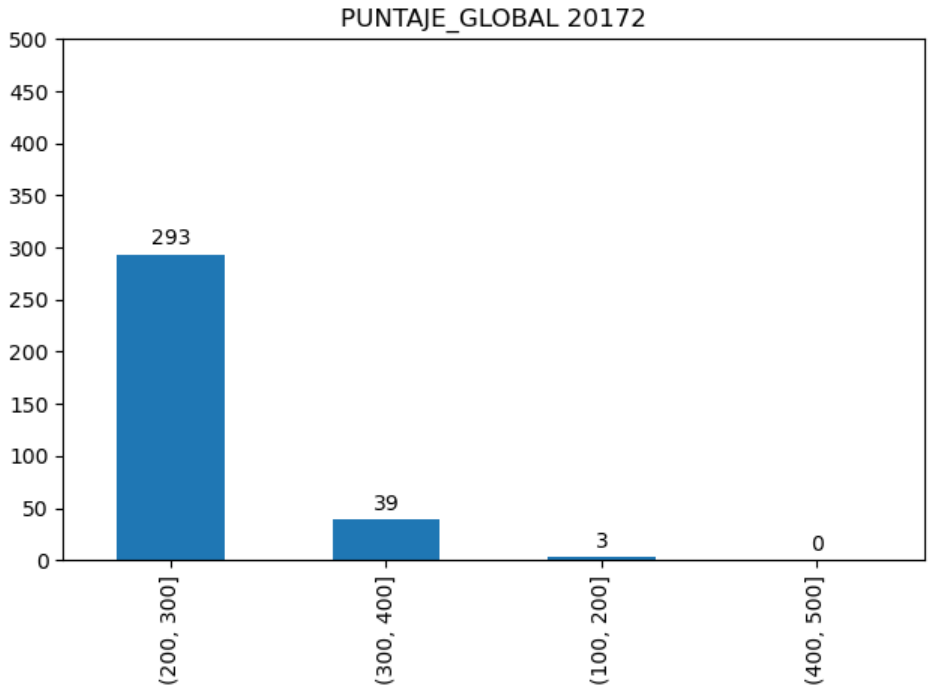


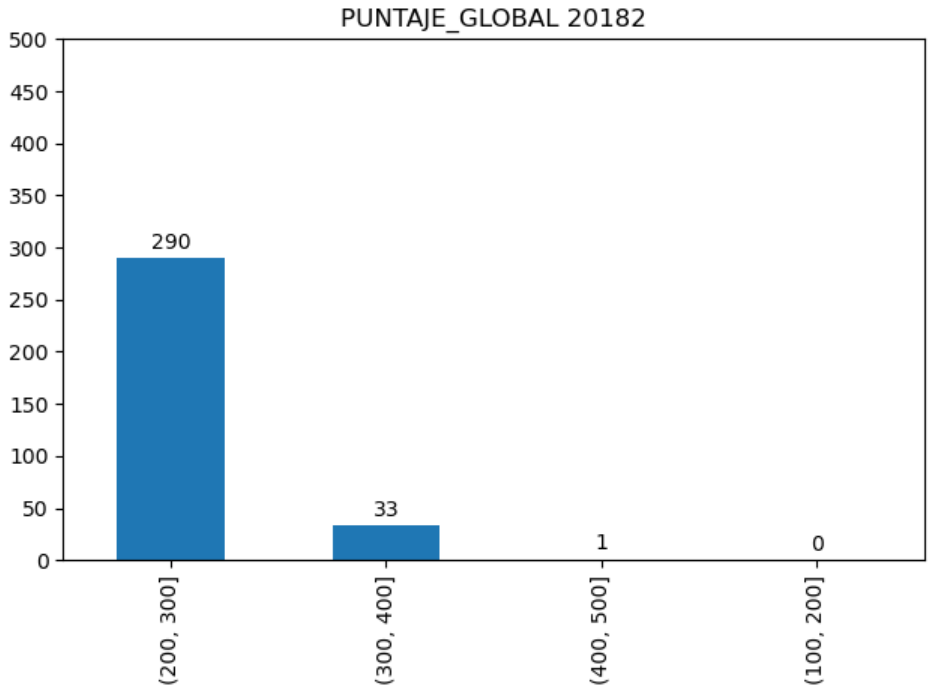
- **Gráfico de frecuencias**
  - → Al ver que la mayoría de los colegios consiguen un puntaje global entre 200 y 300, nos damos cuenta de que la media del puntaje global es acertada, pues está contenida en donde la mayor parte de la población se encuentra concentrada.
  - → También podemos apreciar que hacia los rangos extremos se encuentran muy pocas muestras




PUNTAJE_GLOBAL 20152
500
450
400
350
300
250
200
150
100
50
0
295
32
1
0
(200, 300]
(300, 400]
(100, 200]
(400, 500]


PUNTAJE_GLOBAL 20162
500
450
400
350
300
250
200
150
100
50
0
283
47
2
0
(200, 300]
(300, 400]
(100, 200]
(400, 500]





PARTE II

Pruebas de hipótesis:

- Media poblacional:

Del estudio previo, de forma aleatoria se eligieron 81 establecimientos que en el Período 2018-2 obtuvieron un puntaje global con una media de 262,08 y una desviación estándar de 23,89 ¿Se puede afirmar con un nivel de significancia del 5% que el puntaje global medio real es menor que 280 puntos?

Paso 1

- H0: u >= 280
- H1: u < 280

Paso 2

- Sea α = 0,05

Paso 3

CASO III

- Zc = (262,08 - 280) / (23,89 / sqrt(81))
- Zc = -6,747

Paso 4

- U < u0
- Zc < -Zα
- -6,747 < -1,645

Se rechaza H0, por lo que la hipótesis es cierta, el puntaje medio real es inferior a los

280 puntos en el período del 2018-2

- Proporción poblacional

  Según lo visto, nosotros sospechamos que más del 15% de instituciones educativas obtuvieron un mal desempeño en la prueba de ciencias naturales. Para verificar esto, se eligió una muestra aleatoria del 30% de la cantidad de los datos en el período 2018-2 y vimos que 51 de ellas tuvieron un mal desempeño ¿Qué se puede decir de la sospecha que tenemos con un nivel de significancia del 5%?

  Entiéndase mal desempeño en la prueba como obtener menos de 51 puntos (debajo de la media)

  Paso 1

  - H0: p <= 0.15
  - H1: p > 0.15

  Paso 2

  - Sea α = 0,05

Paso 3

- N = 97
- X = 51
- P gorro = 51 / 97 = 0,525
- Zc = (0,525 – 0,15) / sqrt((0,15 (1-0,15)) / 97) = 10,36

Paso 4

Región crítica dada por: Zc > Zα

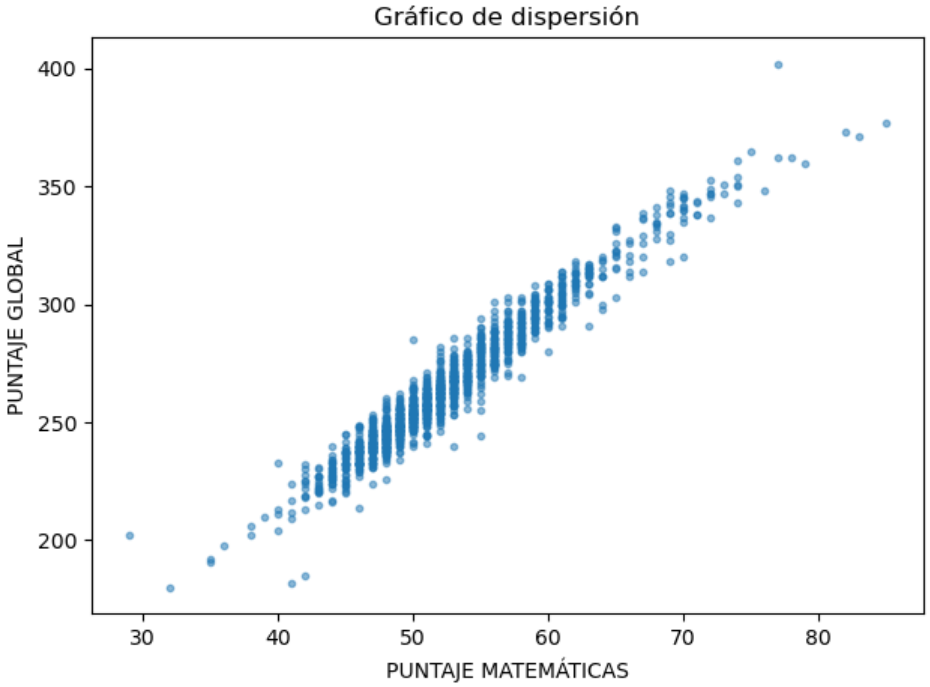
- 10,36 > 1,64

Se rechaza H0, por lo que la hipótesis planteada anteriormente es cierta, más del 15% de las instituciones tuvieron un mal desempeño en ciencias naturales.

Modelo de regresión lineal:
Creemos que el puntaje global está asociado al puntaje de matemáticas obtenido

Paso 1: Diagnóstico de supuestos básicos

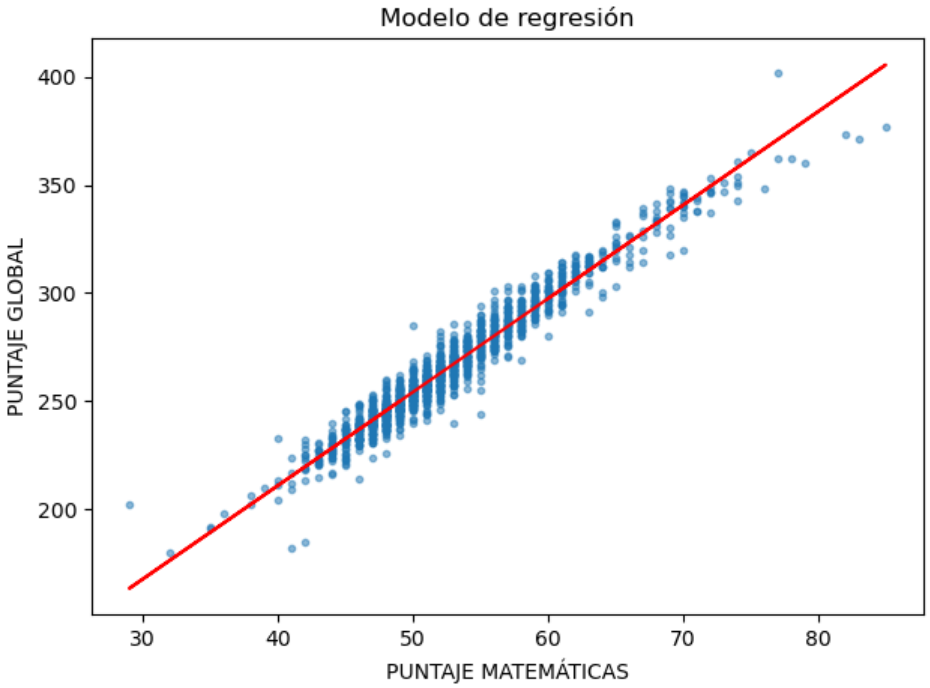
- De Python, obtenemos que r = 0,9724



Paso 2: Estimación de los parámetros del modelo

- Sea y = puntaje global
- Sea x = puntaje matemáticas
- De Python obtenemos que β0 = 37,876
- De Python obtenemos que β1 = 4,325
- Obtenemos la siguiente ecuación:

$$Y = 4{,}325x + 37{,}876$$



Paso 3: Significación estadística del modelo:

- Obtenemos que r2 = 0,945

Significancia F

H0: $\beta 0 = \beta 1 = 0$

H1: uno o más de los parámetros es diferente a cero

- F statistic = 1.881e+06 (obtenido de python)
- Pvalue = 0.0 (obtenido de python)

Con un $\alpha = 0{,}05$

Valor $p < \alpha$

Por lo que rechazamos H0 y decimos que el modelo tiene significancia estadística global

Significancia T

H0: $\beta 1 = 0$

H1: $\beta 1 =! 0$

Con un $\alpha = 0{,}05$

- T statistic = 1371.589
- Pvalue = 0.00

Valor $p < \alpha$

Por lo que rechazamos H0 y decimos que el modelo tiene significancia estadística individual

Paso 4: Test de normalidad

H0: Los datos residuos del modelo se distribuyen normal

H1: Los datos residuos del modelo no se distribuyen normal

Dada la función de Anderson Darling tenemos que

Función = anderson(residual_vals)

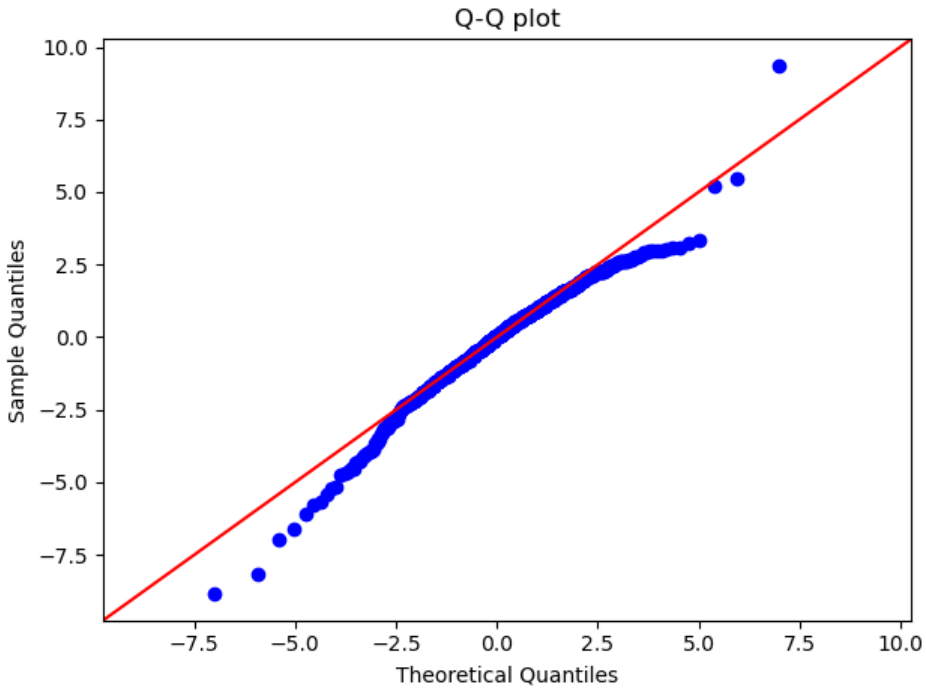
Tiene como salida:

significance_level = array[15, 10, 5, 2.5, 1]

Lo que, dado la documentación del método, indica que tiene una distribución aproximadamente normal

normal/exponenential

15%, 10%, 5%, 2.5%, 1%



Conclusiones generales:

- La cantidad de colegios oficiales que están por encima de la media del puntaje global ha ido creciendo en número, por lo que se puede inferir que hay una mejor calidad de educación en estos establecimientos.
- La desviación estándar es grande en todos los períodos, lo que puede indicar que el aprendizaje no es muy heterogéneo en los colegios.

- La media del puntaje global no ha cambiado mucho con respecto a los años, por lo que se puede decir que no hay ni una disminución, ni un aumento en capacidades.
- El índice de estudiantes matriculados a las pruebas suele ser menor que el de estudiantes presentes en las pruebas.
- Se puede observar que en el 2018-2, la media del puntaje global está por debajo de los 280 puntos, afirmando esto con un nivel de significancia del 5%.
- Vemos que, en ciencias naturales, más del 15% de las instituciones tiene un mal desempeño, esto con un nivel de significancia del 5%.
- Vemos que el modelo de regresión lineal se distribuye aproximadamente normal.
- Vemos que el porcentaje de datos que nuestro modelo de regresión lineal modela es muy bueno, siendo este del 94,5% de los datos.
- Basados en el modelo de regresión lineal podemos predecir que entre más alto sea el puntaje en pruebas de matemáticas, entonces el puntaje global también aumentará.

Referencias:

- ICFES. (2016). Reporte de resultados históricos del examen saber 11, obtenido de https://www.icfes.gov.co/documents/20143/526013/Reporte%20de%20resultados%20historicos%20establecimientos%20educativos.pdf

- Alcaldía de Medellín. (2019). Histórico resultados pruebas saber 11, obtenido de http://medata.gov.co/dataset/historico-resultados-pruebas-saber-11

- Scipy.org. (2020). Scipy.stats.anderson, obtenido de https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.anderson.html