UNIVERSIDAD TECNOLÓGICA DE PANAMÁ
FACULTAD DE INGENIERÍA DE SISTEMAS COMPUTACIONALES
LICENCIATURA EN INGENIERÍA DE SOFTWARE
ESTADÍSTICA CON APOYO INFORMÁTICO

PROYECTO FINAL

PROFESOR

JUAN CASTILLO

PERTENECE A

NALLELY SANCHEZ
8-970-1343

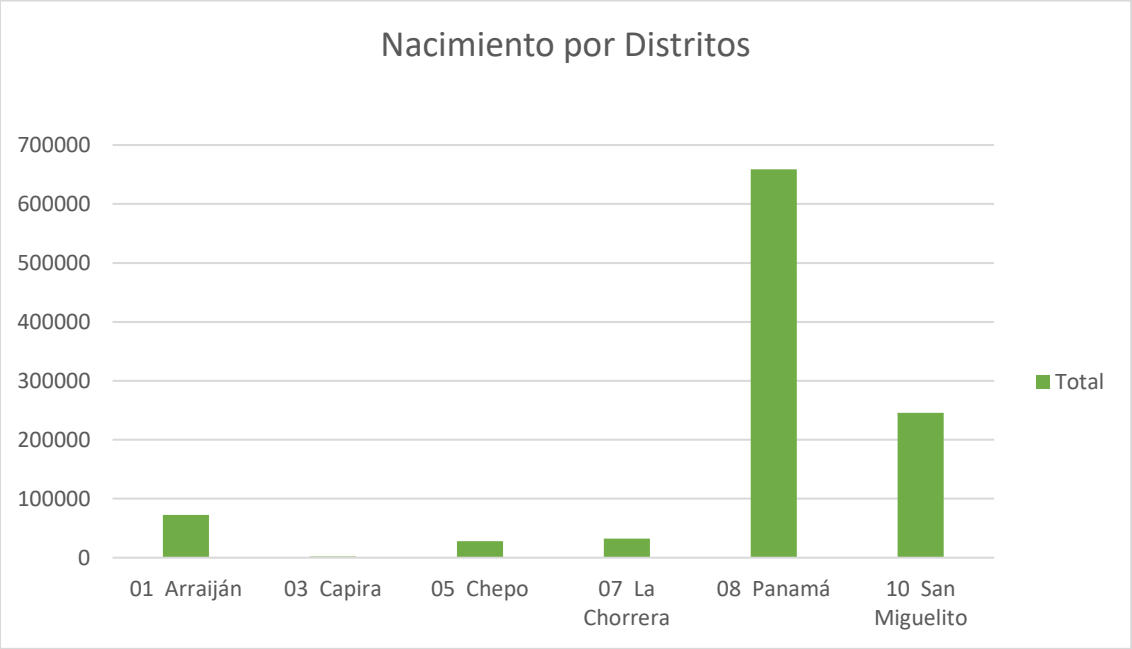
GRUPO 1SF131

1. Nacimientos vivos en la república de Panamá
   Descripción de la aproximación: Curiosidad
   Esta base de datos contiene ID de cada nacimiento, las provincias en donde ocurrieron, el año, el género del nacido, la cantidad de nacimientos, y el estado civil de la mujer que dio a luz.

   Descripción de las variables:
   - Provincias: Texto
   - Año: Numérico
   - Distrito: Texto
   - Mes: Numérico
   - Edad de la madre: Numérico
   - Edad del padre: Numérico
   - Escolaridad de la madre: Texto
   - Estado conyugal: Texto
   - Hijos vivos: Numérico
   - Sexo: Texto
   - Tipo de nacimiento: Texto
   - Área de la madre: Texto

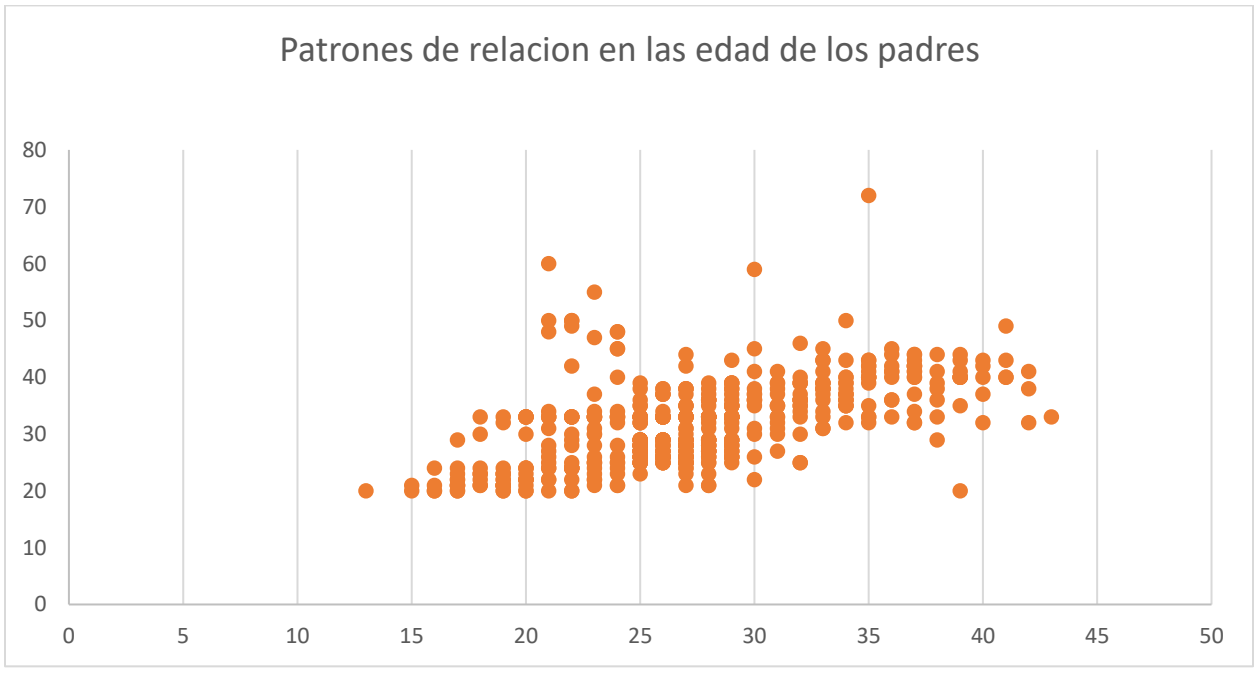
Analítica visual: Gráficos para entender los datos

Podemos observar que el distrito de Panamá es el cual tiene el más alto número de nacimientos vivos

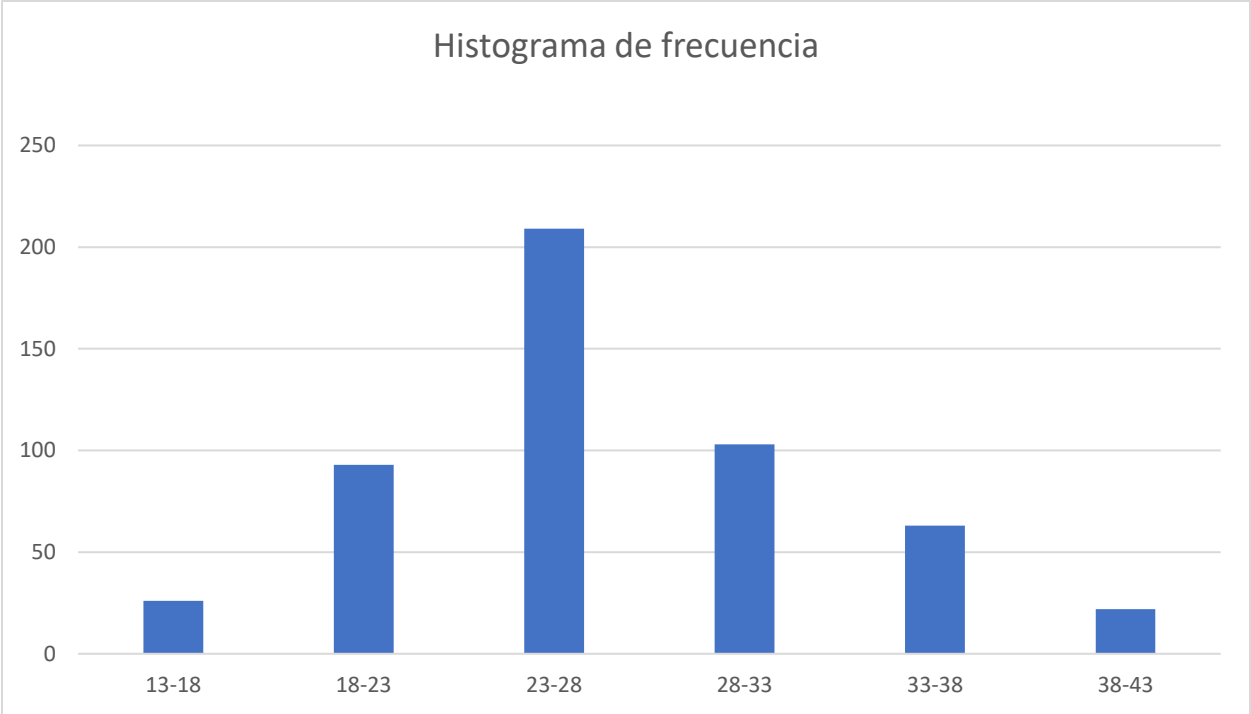


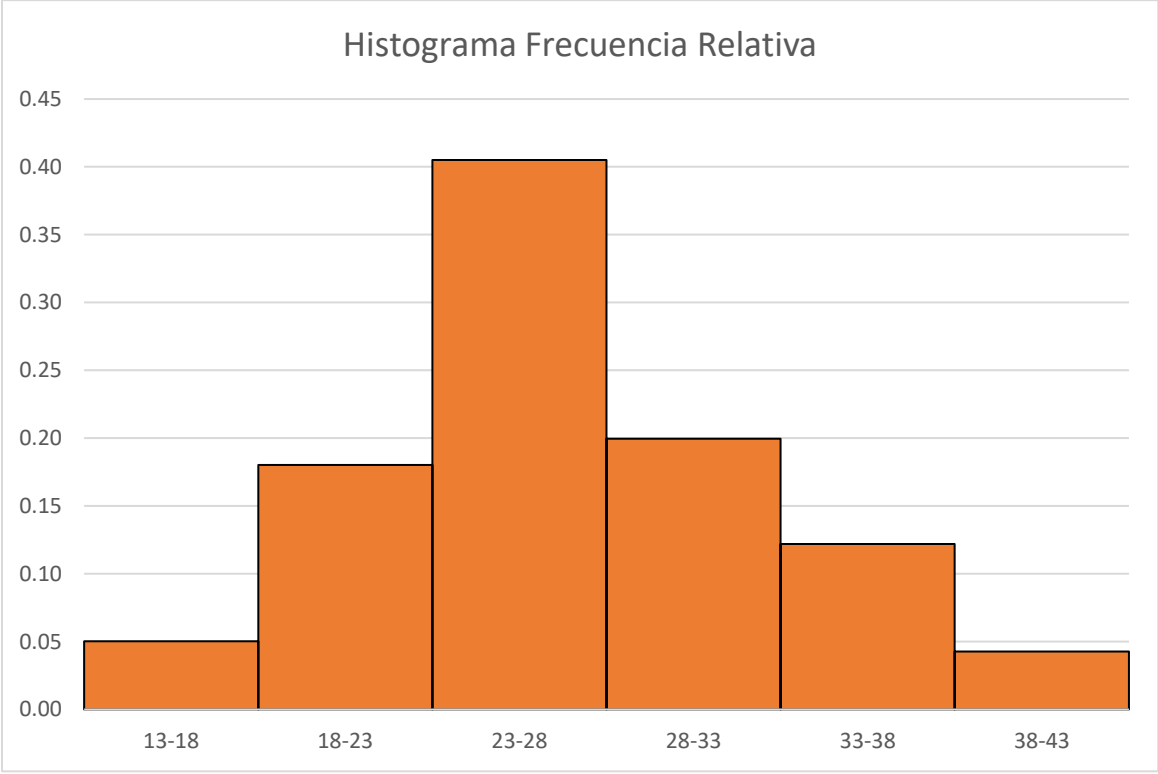
Acá podemos observar una distribución por año la cual nos indica que el 2011 fue el año con el número más alto de nacimientos vivos

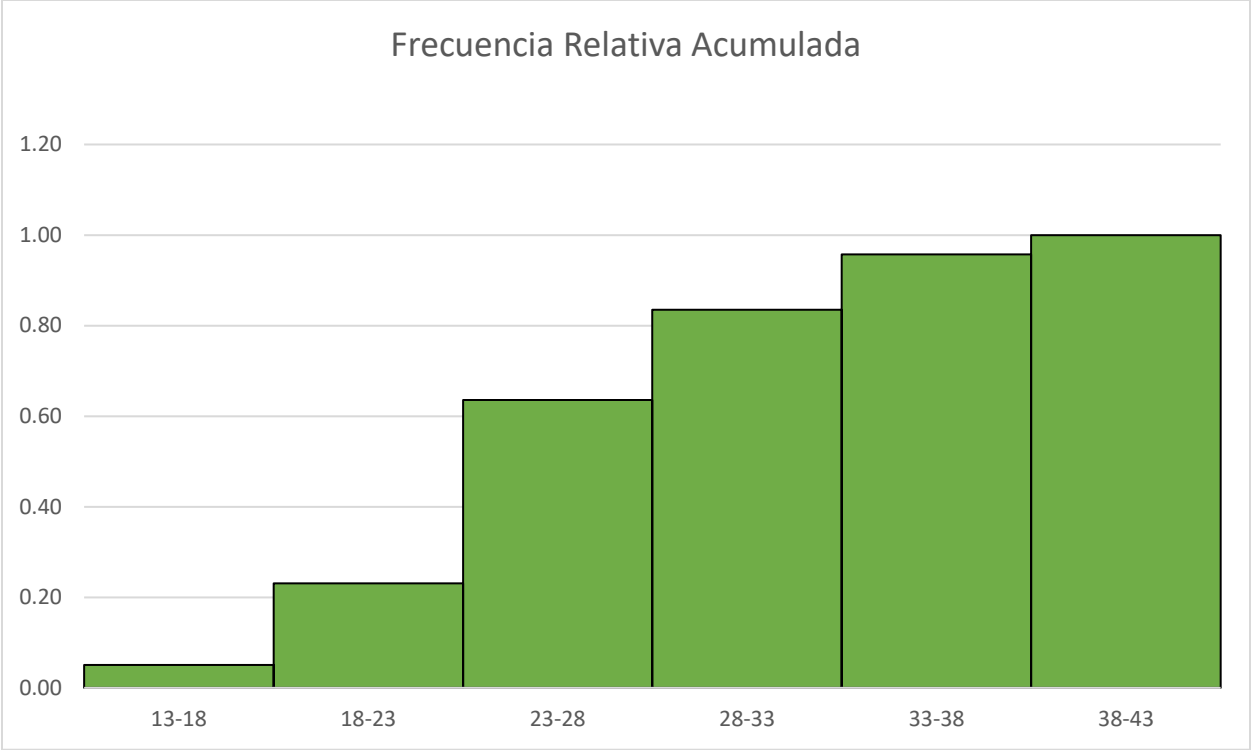
Con este grafico de dispersión podemos observar que la mayor parte de los patrones se encuentran dentro del rango en comparación de las edades de las madres y los padres

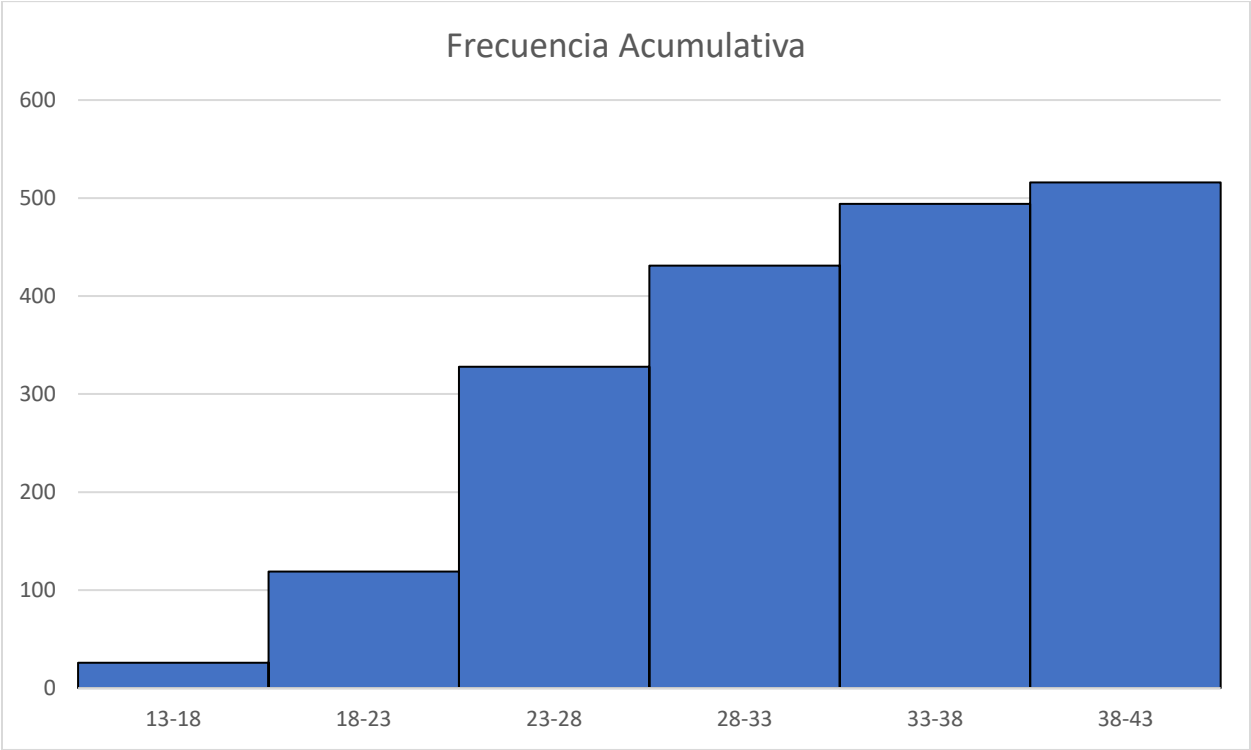


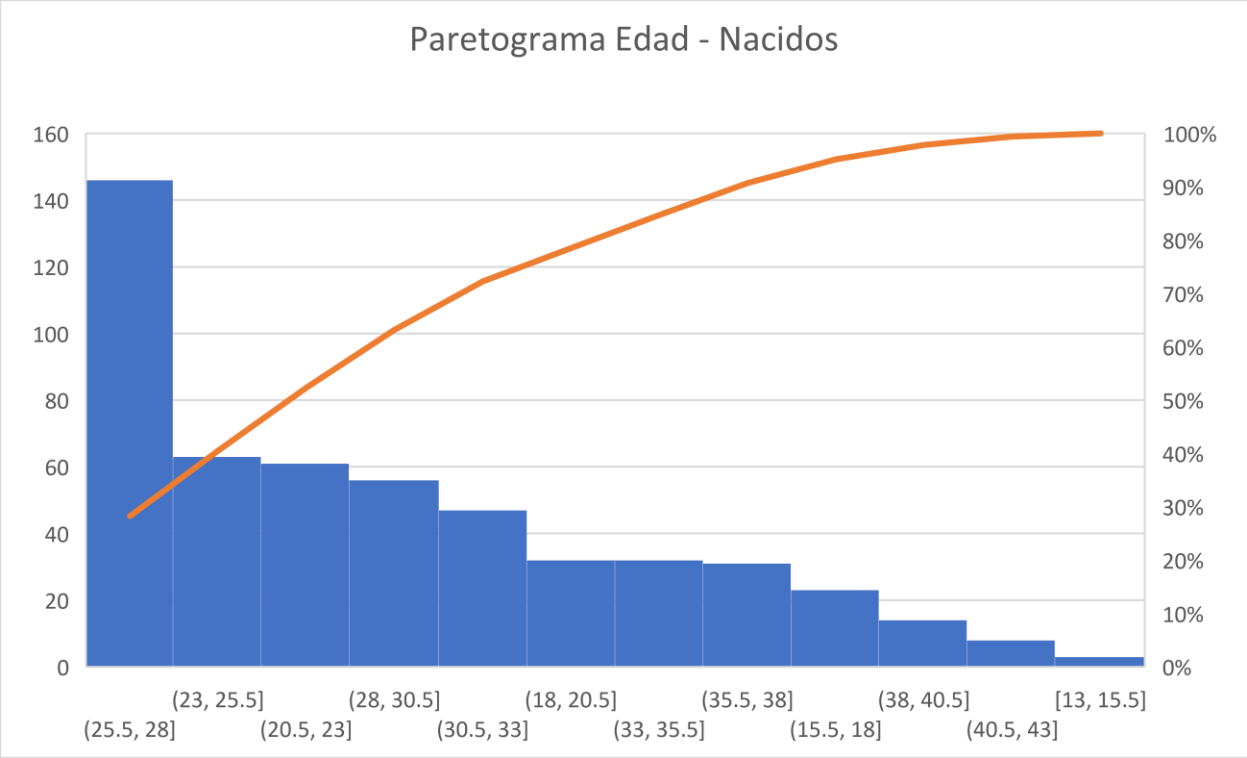
Histogramas de frecuencia a partir de las edades de la madre






Frecuencia Relativa Acumulada
1.20
1.00
0.80
0.60
0.40
0.20
0.00
13-18
18-23
23-28
28-33
33-38
38-43


Frecuencia Acumulativa
600
500
400
300
200
100
0
13-18
18-23
23-28
28-33
33-38
38-43



Media y desviación estándar

Media para las edades de la madre	Varianza	Desviacion estandar
27.50	32.11	5.67

La media representa el valor central de un conjunto de datos, en este caso indica que la edad promedio de las madres en la base de datos es de aproximadamente 27.5 años.

La varianza es aproximadamente 32.1, lo que significa que las edades de las madres varían bastante en relación con la media.

La desviación estándar es aproximadamente 5.7, lo que sugiere que las edades de las madres se encuentran bastante dispersas en relación con la edad promedio.

Media por edad del padre	Varianza	Desviacion estandar
31.80	52.73	7.26

La media de las edades de los padres es de 31.80 años. Esto significa que, en promedio, los padres tienen aproximadamente 31 – 32 años.

La varianza de las edades de los padres es de 52.73. La varianza es una medida de la dispersión o la variabilidad de los datos. Un valor de varianza más alto indica que las edades de los padres están más dispersas alrededor de la media.

La desviación estándar de las edades de los padres es de 7.26 años. La desviación estándar es otra medida de la dispersión de los datos. Una desviación estándar más alta indica que las edades de los padres tienden a estar más alejadas de la media.

Análisis de Correlación

Coeficiente de correlación (Multiple R): El coeficiente de correlación Multiple R (0.2552) indica la fuerza y la dirección de la relación lineal entre la variable independiente "5" y el resultado. Un valor de 0.2552 sugiere que existe una correlación positiva débil entre estas variables.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	24.9874813	0.48402218	51.62466	2.7E-205	24.03657178	25.93839081	24.03657178	25.93839081
5	1.119208268	0.187197639	5.978752	4.22E-09	0.751439966	1.486976569	0.751439966	1.486976569

Análisis de Regresión Lineal

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1079.292737	1079.293	35.74548	4.21541E-09
Residual	513	15489.43348	30.19383		
Total	514	16568.72621			

Coeficientes: En el modelo de regresión, se han obtenido dos coeficientes. El coeficiente del término "Intercept" (24.987) representa el valor estimado del resultado (variable dependiente) cuando la variable independiente (5) es igual a cero. El coeficiente asociado a la variable independiente "5" (1.119) indica cómo cambia el valor estimado del resultado cuando la variable "5" aumenta en una unidad.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	24.9874813	0.48402218	51.62466	2.7E-205	24.03657178	25.93839081	24.03657178	25.93839081
5	1.119208268	0.187197639	5.978752	4.22E-09	0.751439966	1.486976569	0.751439966	1.486976569

Significancia: Los valores "p" asociados a los coeficientes son extremadamente pequeños (4.21541E-09), lo que indica que ambos coeficientes son estadísticamente significativos. En otras palabras, hay evidencia sólida de que la variable independiente "5" tiene un impacto significativo en el resultado.

Regression Statistics	
Multiple R	0.25522609
R Square	0.065140357
Adjusted R Square	0.063318018

Standard Error	5.494891031
Observations	515

R Square: El coeficiente de determinación R Square (0.0651) representa la proporción de la variabilidad total del resultado que se puede explicar por la variable independiente "5". En este caso, alrededor del 6.51% de la variabilidad del resultado puede explicarse por la variable "5". Esto sugiere que, aunque hay una relación significativa entre las variables, la variable independiente "5" solo explica una pequeña parte de la variabilidad del resultado.

2. Embarazo:

Descripción de la aproximación: Curiosidad

Descripción de variables:

Provincia: Texto

Esta variable representa la provincia a la que pertenece la información relacionada con los nacimientos.

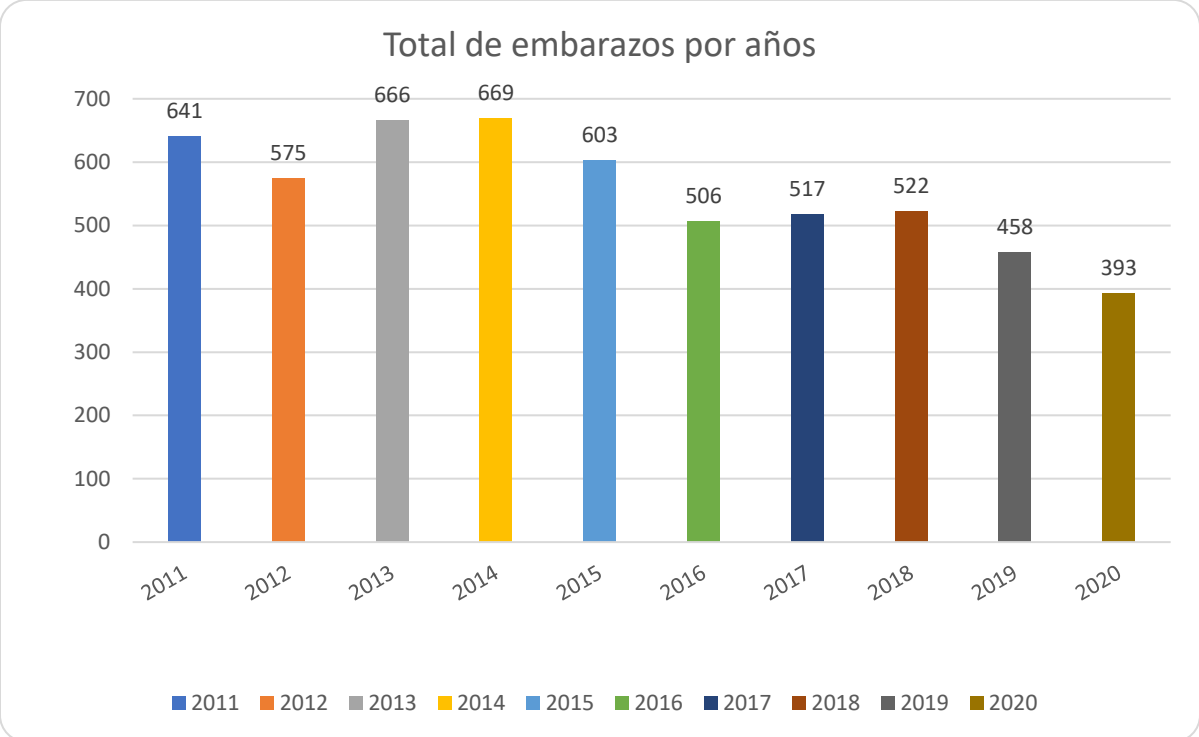
Año: Numérico

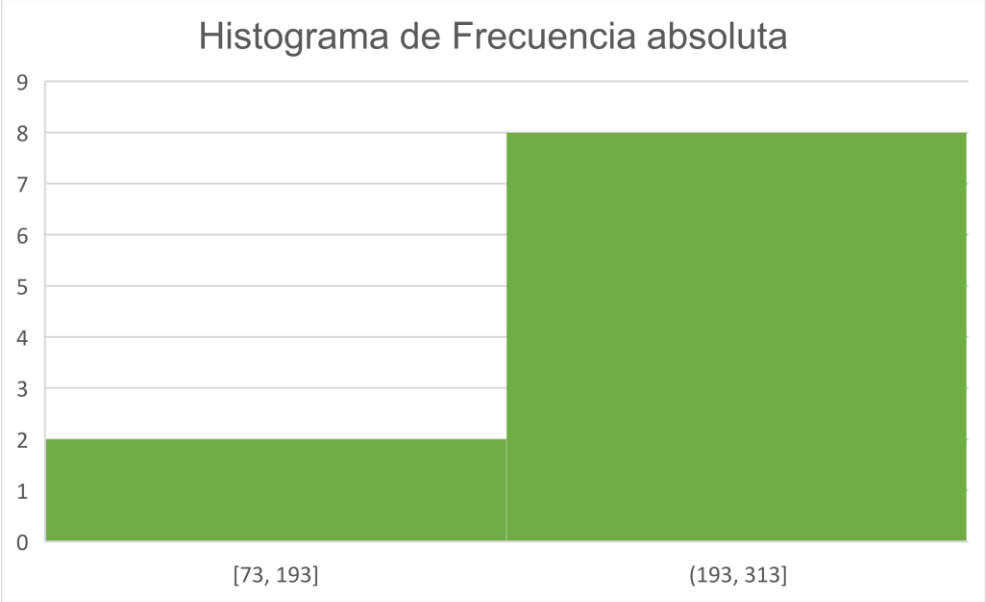
Esta variable indica el año en el que se registraron los nacimientos. Es un dato temporal que proporciona información sobre el período al que pertenece cada registro de nacimiento.

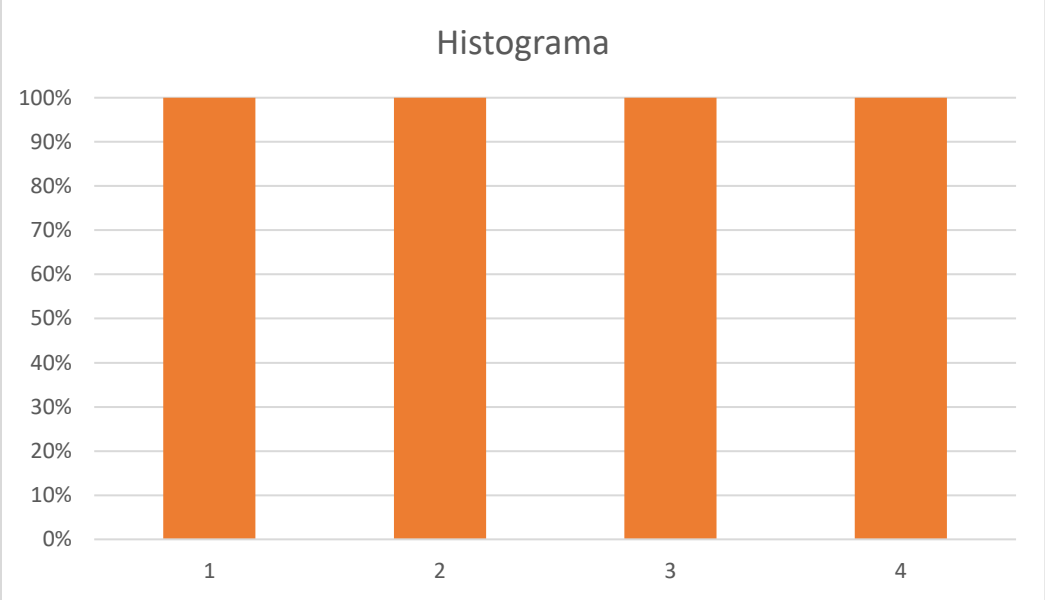
Nacimientos: Numérico

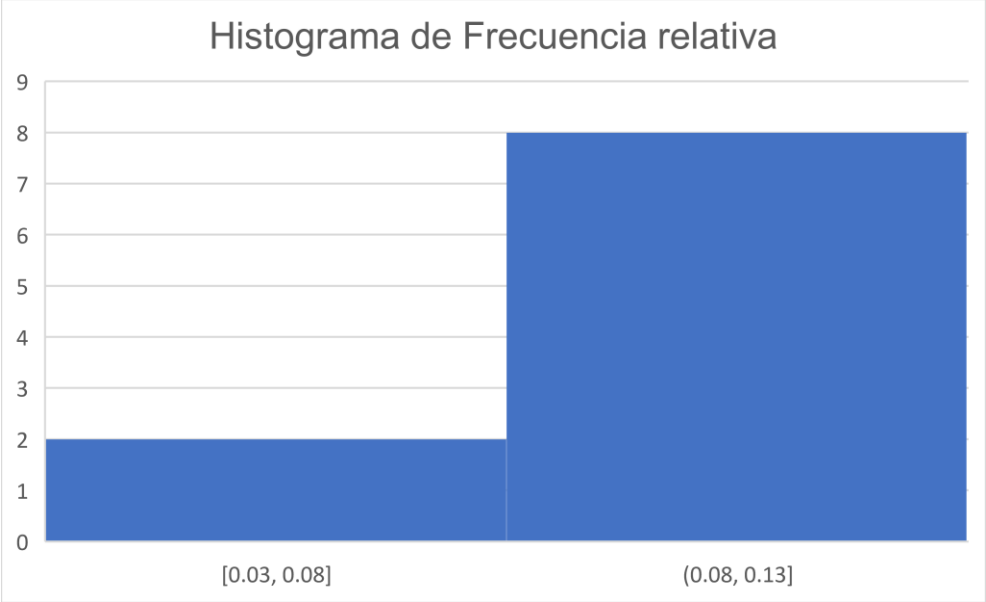
Esta variable indica el número de nacimientos registrados en un año específico en una determinada provincia. Es una variable numérica que nos muestra la cantidad de bebés nacidos en ese período y ubicación particular.

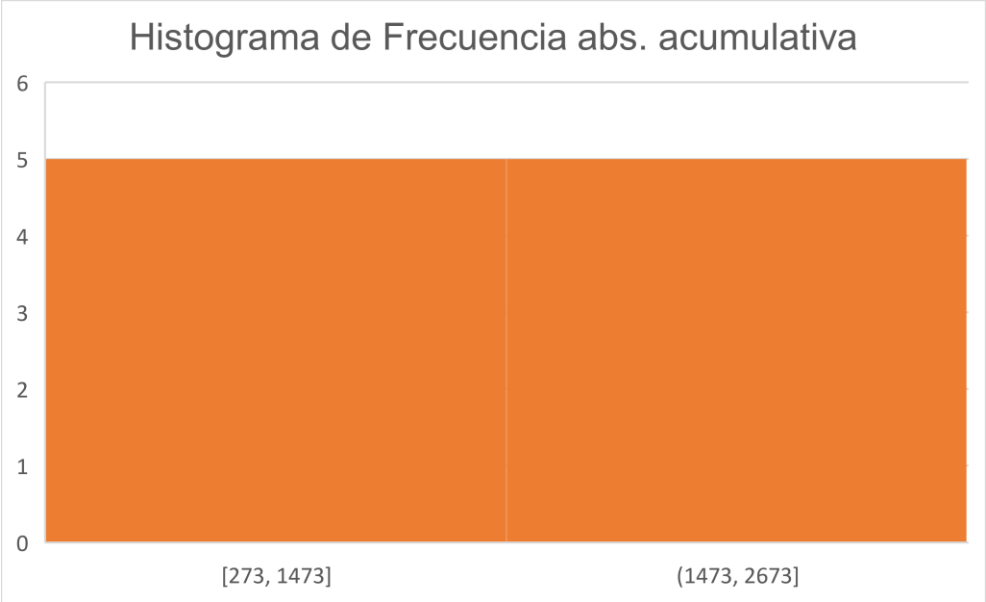
Análisis visual




Histograma de Frecuencia absoluta
9
8
7
6
5
4
3
2
1
0
[73, 193]
(193, 313]


Histograma
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4


Histograma de Frecuencia relativa
9
8
7
6
5
4
3
2
1
0
[0.03, 0.08]
(0.08, 0.13]


Histograma de Frecuencia abs. acumulativa
6
5
4
3
2
1
0
[273, 1473]
(1473, 2673]

Análisis de Correlación

Intervalo (10-14)	Fr. absoluta	Fr. ab. acum.	Fr. relativa	Fr. rel. acum.	Suma de los intervalos	Media por intervalo	Análisis exponencial	Covarianza	Desviación estándar	Correlación
2010-2011	48	48	17.65	17.65	1249	26.02	1.9985E+11	0	1698.47049	0
2012-2013	47	95	17.28	34.93	1241	26.40	2.9324E+11	0	1688.57099	0
2014-2015	51	146	18.75	53.68	1272	24.94	6.7891E+10	0	1726.75476	0
2016-2017	50	196	18.38	72.06	1023	20.46	768537564	0	1376.0298	0
2018-2019	51	247	18.75	90.81	980	19.22	221445324	0	1313.8044	0
2020	25	272	9.19	100.00	393	15.72	6715977.86	0	520.430591	0
Total	272	1004	100	369.12						

- Distribución de embarazos: El análisis muestra la distribución de embarazos agrupados en intervalos de años, desde 2010 hasta 2020. Cada intervalo contiene la cantidad de embarazos ocurridos en ese período.

- Frecuencias absolutas y relativas: Los resultados muestran la cantidad de embarazos en cada intervalo (frecuencia absoluta) y el porcentaje de embarazos en relación con el total (frecuencia relativa). Por ejemplo, en el intervalo 2010-2011, hubo 48 embarazos, lo que representa el 17.65% del total de embarazos.

- Frecuencias acumuladas: La columna "Fr. ab. acum." muestra la frecuencia absoluta acumulada, es decir, la suma de las frecuencias absolutas hasta el intervalo actual. La columna "Fr. rel. acum." muestra la frecuencia relativa acumulada, que representa la suma de las frecuencias relativas hasta el intervalo actual.

- Media por intervalo: La columna "Media por intervalo" representa el promedio de embarazos en cada intervalo. Por ejemplo, en el intervalo 2010-2011, el promedio de embarazos fue de aproximadamente 26.02.

- Análisis exponencial, Covarianza y Correlación: Los valores de "Análisis exponencial", "Covarianza" y "Correlación" son iguales a cero, lo que sugiere que no hay una relación exponencial, covarianza ni correlación significativa entre los intervalos de años y la cantidad de embarazos.

Análisis de Regresión Lineal

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.14680869
Coeficiente de determinación R^2	0.021552792
R^2 ajustado	-0.087163565
Error típico	1.12475022
Observaciones	11

- Coeficiente de correlación múltiple: El coeficiente de correlación múltiple es 0.1468. Este valor indica la fuerza y la dirección de la relación lineal entre las variables analizadas (embarazos y años). Un coeficiente de correlación cercano a 1 indica una correlación positiva fuerte, mientras que un valor cercano a -1 indica una correlación negativa fuerte. En este caso, el valor 0.1468 sugiere una correlación positiva débil entre los embarazos y los años, lo que significa que hay una tendencia de aumento de embarazos a medida que avanzan los años, pero esta relación es relativamente débil.

- Coeficiente de determinación R^2: El coeficiente de determinación R^2 es 0.0216, lo que indica que aproximadamente el 2.16% de la variabilidad de los embarazos puede explicarse por los años. En otras palabras, solo un pequeño porcentaje de la variación en la cantidad de embarazos se puede atribuir a la variable de años.

- R^2 ajustado: El valor de R^2 ajustado es -0.0872. El R^2 ajustado tiene en cuenta el número de variables en el modelo y penaliza el uso excesivo de variables para evitar sobreajuste. Un valor negativo en el R^2 ajustado puede indicar que el modelo no se ajusta bien a los datos y que el número de variables puede no ser adecuado.

- Error típico: El error típico es 1.1247. Este valor representa la variabilidad de los datos con respecto al ajuste del modelo de regresión. Un error típico más bajo indica que el modelo se ajusta mejor a los datos.

- Observaciones: El análisis se basó en 11 observaciones, es decir, datos de 11 años diferentes.

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	0.25079612	0.25079612	0.19824792	0.66665533
Residuos	9	11.3855675	1.26506306		
Total	10	11.6363636			

Intercepción	25.81844251	2.27196504	11.363926	1.22282E-06	20.6789005	30.9579845	20.6789005	30.9579845
Variable X 1	-0.00178676	0.00401293	-0.4452504	0.666655328	-0.01086464	0.00729112	-0.01086464	0.00729112

- Grados de libertad: El análisis de varianza muestra que hay un total de 10 grados de libertad, que se distribuyen en 1 grado de libertad para la regresión y 9 grados de libertad para los residuos.

- Suma de cuadrados: La suma de cuadrados para la regresión es 0.2508, mientras que la suma de cuadrados para los residuos es 11.3856. La suma de cuadrados mide la variabilidad total de los datos y se descompone en la variabilidad explicada por la regresión y la variabilidad no explicada (residuos).

- Promedio de los cuadrados: El promedio de los cuadrados es el cociente de la suma de cuadrados entre los grados de libertad correspondientes. Para la regresión, el promedio de los cuadrados es 0.2508, y para los residuos, es 1.2651. El promedio de los cuadrados representa la variabilidad promedio en los datos para cada fuente de variación.

- F y Valor crítico de F: El valor de F obtenido para la regresión es 0.1982. El valor crítico de F se utiliza para comparar el valor de F obtenido con un valor crítico de referencia y determinar si la regresión es significativa o no. En este caso, el valor de F es menor que el valor crítico de referencia (0.6667), lo que sugiere que la regresión no es significativa. En otras palabras, no hay suficiente evidencia para concluir que la regresión explica de manera significativa la variabilidad en los datos.

Análisis de los residuales

Observación	*Pronóstico para Y*	*Residuos*
1	24.73209248	-0.73209248
2	24.6731294	-0.6731294
3	24.79105555	-1.79105555
4	24.6284604	-0.6284604
5	24.62310012	1.37689988
6	24.74102628	1.25897372
7	24.91434199	-0.91434199
8	24.89468763	1.10531237
9	24.88575383	1.11424617
10	25.00010646	-0.00010646
11	25.11624586	-0.11624586

Resultados de datos de probabilidad

Percentil	*Y*
4.545454545	23
13.63636364	24

22.72727273	24
31.81818182	24
40.90909091	24
50	25
59.09090909	25
68.18181818	26
77.27272727	26
86.36363636	26
95.45454545	26

- Pronóstico y Residuos: El pronóstico para Y representa los valores predichos por el modelo para cada observación. Los residuos son las diferencias entre los valores observados y los valores pronosticados. Por ejemplo, en la primera observación, el pronóstico para Y es 24.7321, mientras que el valor observado es 25. El residuo es la diferencia entre estos dos valores, que es -0.7321.

- Percentiles y Valores Y: Se han calculado los percentiles para los residuos, lo que nos permite evaluar cómo se distribuyen los residuos en relación con los valores de Y. Por ejemplo, el percentil 4.55 indica que el 4.55% de los residuos son menores o iguales a -0.7321 (el residuo de la primera observación). Además, el percentil 95.45 indica que el 95.45% de los residuos son menores o iguales a -0.1162 (el residuo de la undécima observación).

Prueba t para medias de dos muestras emparejadas

	Variable 1	*Variable 2*
Media	559.8181818	24.8181818
Varianza	7855.763636	1.16363636
Observaciones	11	11
Coeficiente de correlación de Pearson	-0.14680869	
Diferencia hipotética de las medias	0	
Grados de libertad	10	
Estadístico t	19.98247488	
P(T<=t) una cola	1.08227E-09	
Valor crítico de t (una cola)	1.812461123	
P(T<=t) dos colas	2.16453E-09	
Valor crítico de t (dos colas)	2.228138852	

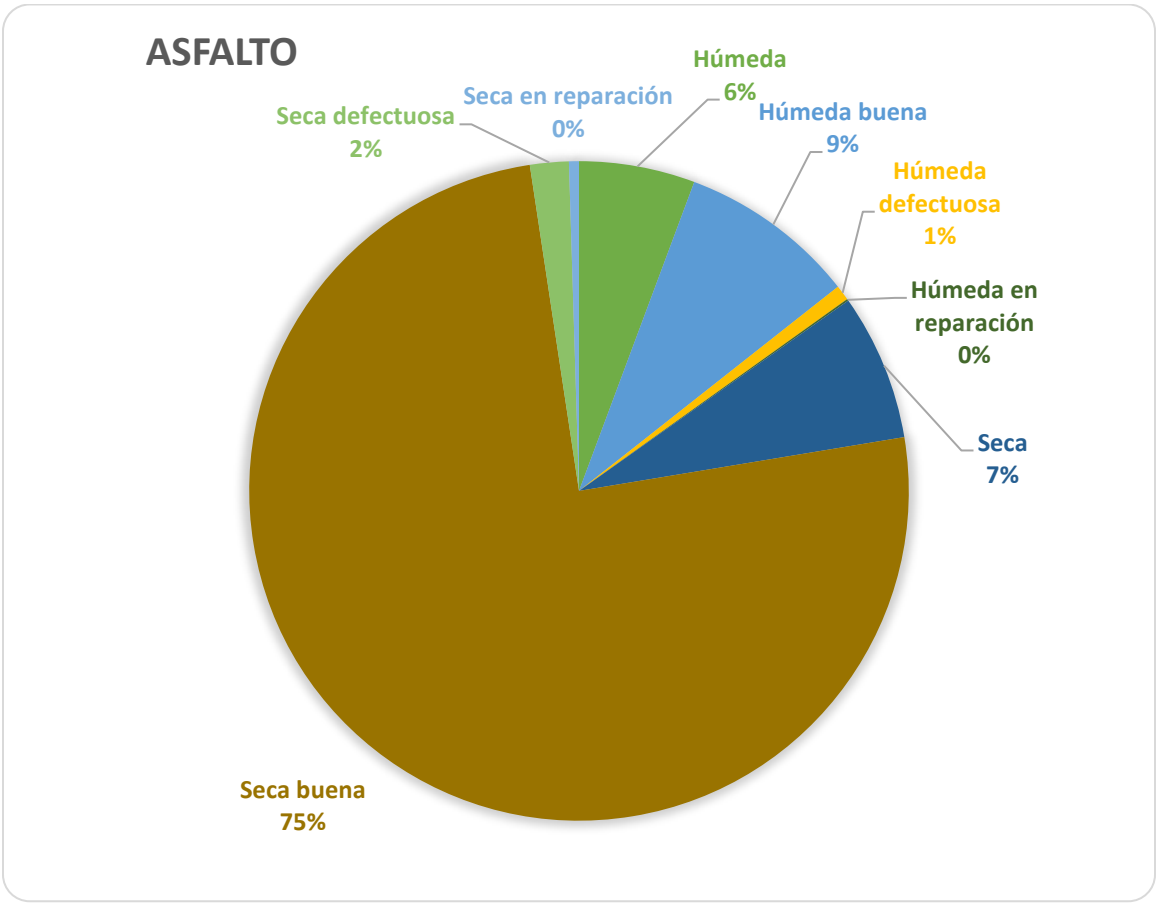
Prueba F para varianzas de dos muestras

	Variable 1	*Variable 2*
Media	559.8181818	24.8181818
Varianza	7855.763636	1.16363636
Observaciones	11	11
Grados de libertad	10	10

F	6751.046875
P(F<=f) una cola	8.97385E-18
Valor crítico para F (una cola)	2.978237016

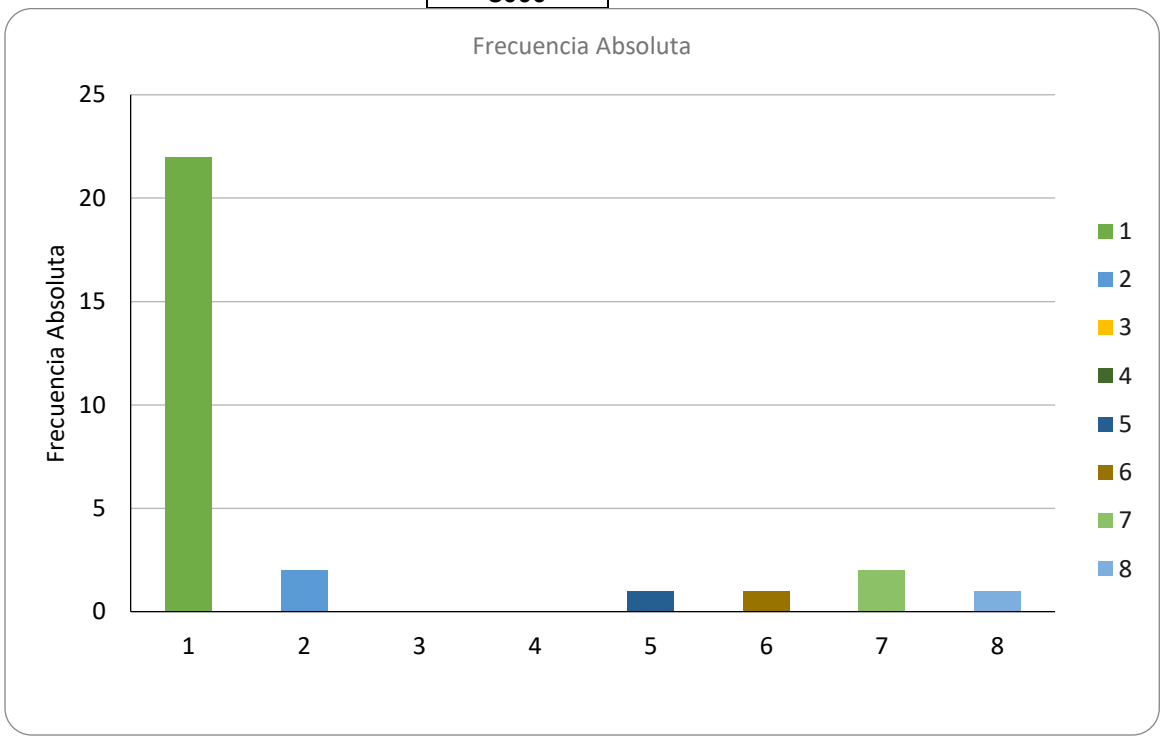
3. Accidentes de tránsito:
   Descripción de la aproximación: Curiosidad
   Descripción de variables:
   - Localización: Abarca las provincias de Panamá. Es de tipo texto.
   - Corregimiento: Abarca los corregimientos de Panamá. Es de tipo texto.
   - Distrito: Abarca los distritos de Panamá. Es de tipo texto.
   - Mes: Meses del año de tipo texto.
   - Área: Categorizado en Urbana o Rural. Es de tipo texto.
   - Calle: Abarca las carreteras donde hayan ocurrido estos accidentes.
   - Clase de Accidente: Especifica si es colisión o vuelco. Es de tipo texto.
   - Clase vía: Especifica si es concreto o asfalto. Es de tipo texto.

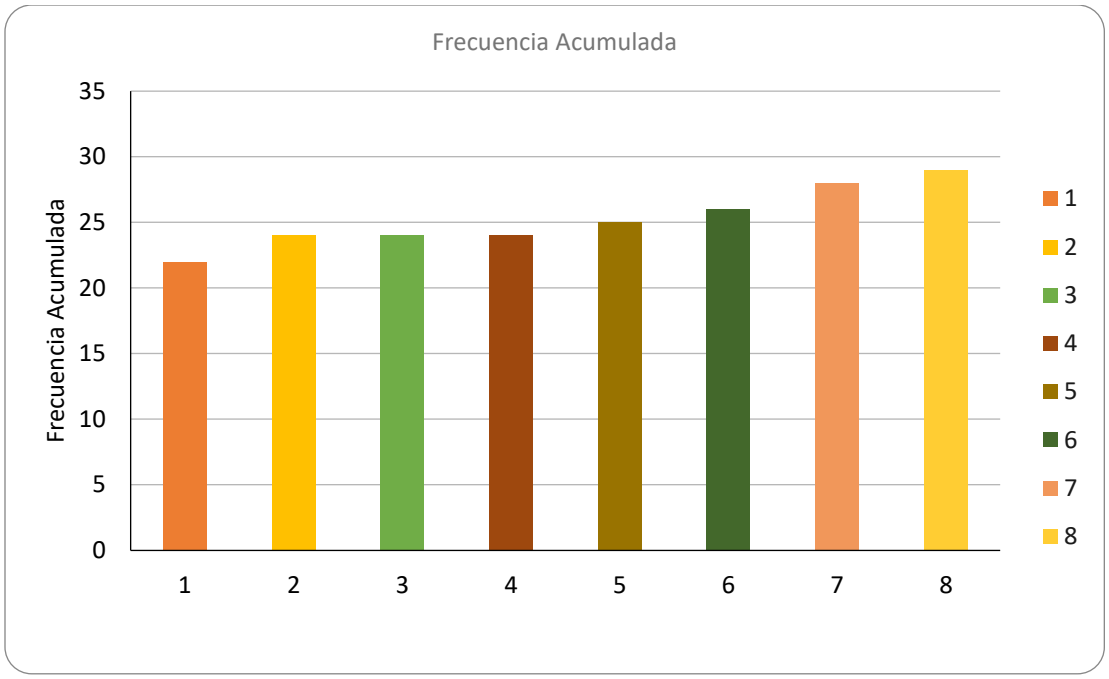
   Histogramas:
   Podemos observar una clasificación del estado del asfalto al momento del accidente.
   Siendo el estado de “seco y buena” la característica que más sobresale dentro de este grupo.

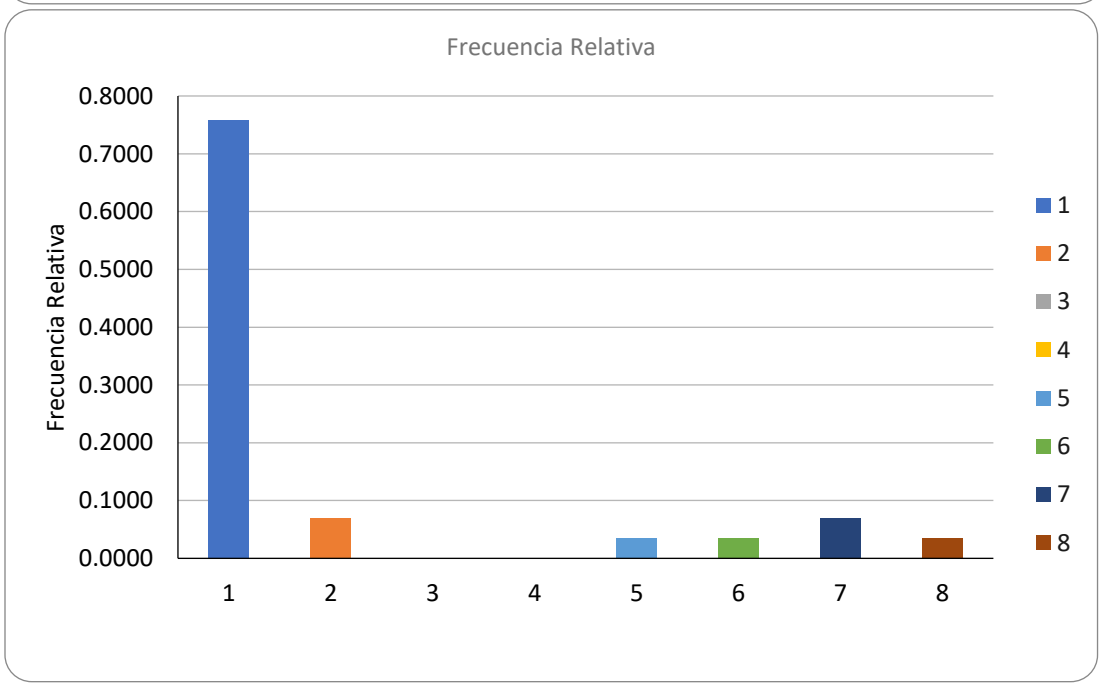


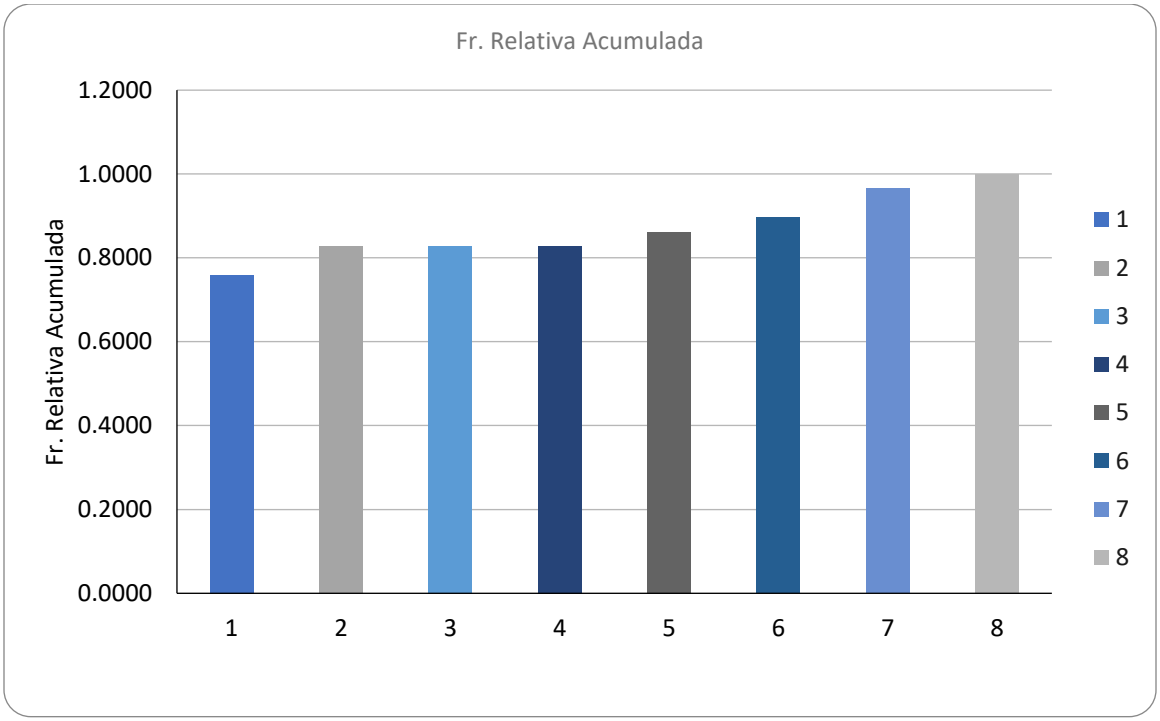
A partir de nuestra tabla de frecuencia logramos obtener sus respectivos histogramas. Teniendo en consideración el intervalo del rango seleccionado.

Intervalo
0 - 299
300 - 599
600 - 899
900 - 1199
1200 - 1499
1500 - 1799
1800 - 2099
mas de 3000




Frecuencia Acumulada
Frecuencia Acumulada
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8


Frecuencia Relativa
Frecuencia Relativa
0.8000
0.7000
0.6000
0.5000
0.4000
0.3000
0.2000
0.1000
0.0000
1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8



Media (alfalto)	Desviación estandar (asfalto)	Varianza (asfalto)
2891.88	5530.92	34961181.27
Media (concreto)	**Desviación estandar (concreto)**	**Varianza (concreto)**
389.13	658.48	495539.84
Media (grava)	**Desviación estandar (grava)**	**Varianza (grava)**
4.38	4.20	22.00
Media (tierra)	**Desviación estandar (tierra)**	**Varianza (tierra)**
42.38	31.12	1106.84

Asfalto:

- La media de accidentes en vías de asfalto es de aproximadamente 2891.88. Esto nos indica que, en promedio, se registran alrededor de 2892 accidentes en vías de asfalto en el área estudiada.
- La desviación estándar de 5530.92 nos indica que los valores de accidentes en vías de asfalto varían bastante alrededor de la media. Esto podría sugerir que hay algunas áreas con una mayor concentración de accidentes en vías de asfalto.

- La varianza de 34,961,181.27 indica la dispersión de los datos en vías de asfalto. Es un valor alto, lo que sugiere una mayor variabilidad en el número de accidentes registrados.
- Concreto:

- La media de accidentes en vías de concreto es de aproximadamente 389.13. Esto indica que hay una cantidad significativamente menor de accidentes en vías de concreto en comparación con las de asfalto.
- La desviación estándar de 658.48 nos indica que los valores de accidentes en vías de concreto también varían, aunque en menor medida que en las vías de asfalto.
- La varianza de 495,539.84 indica una menor dispersión de datos en vías de concreto en comparación con las de asfalto.

Grava:

- La media de accidentes en vías de grava es de aproximadamente 4.38. Esto sugiere que se registran pocos accidentes en este tipo de vías.
- La desviación estándar de 4.20 es bastante cercana a la media, lo que indica que hay poca variabilidad en los datos en vías de grava.
- La varianza de 22.00 indica que los datos están bastante agrupados alrededor de la media en vías de grava.

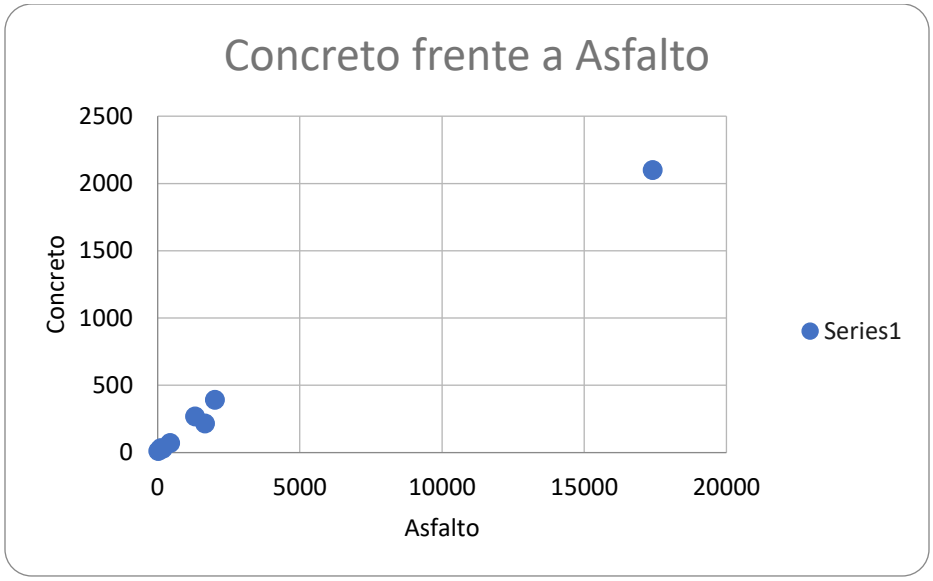
Tierra:

- La media de accidentes en vías de tierra es de aproximadamente 42.38. Esto nos muestra que hay más accidentes en vías de tierra en comparación con las de grava, pero mucho menos que en las de asfalto y concreto.
- La desviación estándar de 31.12 nos indica que los valores de accidentes en vías de tierra también varían, pero en menor medida que en vías de asfalto y concreto.
- La varianza de 1,106.84 indica una dispersión moderada de los datos en vías de tierra.

Análisis de Correlación.

	Asfalto	Concreto
Húmeda	1315	268
Húmeda buena	2010	392
Húmeda defectuosa	173	28
Húmeda en reparación	20	10
Seca	1662	215
Seca buena	17403	2098
Seca defectuosa	439	70
Seca en reparación	113	32

	Asfalto	*Concreto*
Asfalto	1	
Concreto	0.996915058	1



Correlación entre Asfalto y Accidentes:

El coeficiente de correlación entre el tipo de pavimento "Asfalto" y la proporción de accidentes en vías de asfalto es de 1. Esto significa que hay una correlación positiva perfecta entre el tipo de pavimento "Asfalto" y la proporción de accidentes en vías de asfalto. En otras palabras, en todas las situaciones en las que se registra un accidente, el pavimento es de tipo "Asfalto".

Correlación entre Concreto y Accidentes:

El coeficiente de correlación entre el tipo de pavimento "Concreto" y la proporción de accidentes en vías de concreto es de aproximadamente 0.997. Esto también indica una correlación positiva fuerte entre el tipo de pavimento "Concreto" y la proporción de accidentes en vías de concreto. En otras palabras, en la mayoría de las situaciones en las que se registra un accidente, el pavimento es de tipo "Concreto".

Análisis de Regresión Lineal

Regression Statistics	
Multiple R	0.996915058
R Square	0.993839633
Adjusted R Square	0.992812905
Standard Error	501.267719
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	243220652.9	243220652.9	967.9679438	7.32277E-08
Residual	6	1507615.957	251269.3261		
Total	7	244728268.9			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
Intercept	-366.4998943	205.8567739	-1.780363537	0.125313568	-870.2132731	137.2134846	-870.2132731	137.2134846
X Variable 1	8.373594332	0.269141972	31.11218321	7.32277E-08	7.715027653	9.03216101	7.715027653	9.03216101

Coeficiente de Determinación (R Cuadrado):

El coeficiente de determinación (R cuadrado) es de aproximadamente 0.994. Esto significa que aproximadamente el 99.4% de la variabilidad en la proporción de accidentes puede ser explicada por la relación entre el tipo de pavimento y la proporción de accidentes en cada tipo de vía. Es un valor muy alto, lo que indica que la regresión es altamente significativa y que el modelo explica muy bien la variabilidad en los datos.

Coeficiente de Correlación (R):

El coeficiente de correlación (R) es de aproximadamente 0.997. Esto indica una correlación positiva fuerte entre el tipo de pavimento y la proporción de accidentes en cada tipo de vía. En otras palabras, existe una relación significativa entre el tipo de pavimento y la frecuencia de accidentes.

Coeficientes de Regresión:

El coeficiente de regresión para la variable "X Variable 1" (posiblemente la variable independiente que representa el tipo de pavimento) es de aproximadamente 8.374. Esto sugiere que, en promedio, el cambio en la proporción de accidentes está relacionado con un cambio de 8.374 unidades en el tipo de pavimento.

P-Values:

Los valores p para ambos coeficientes son extremadamente bajos (7.32277E-08). Esto indica que ambos coeficientes son estadísticamente significativos, lo que significa que existe una relación significativa entre el tipo de pavimento y la proporción de accidentes.

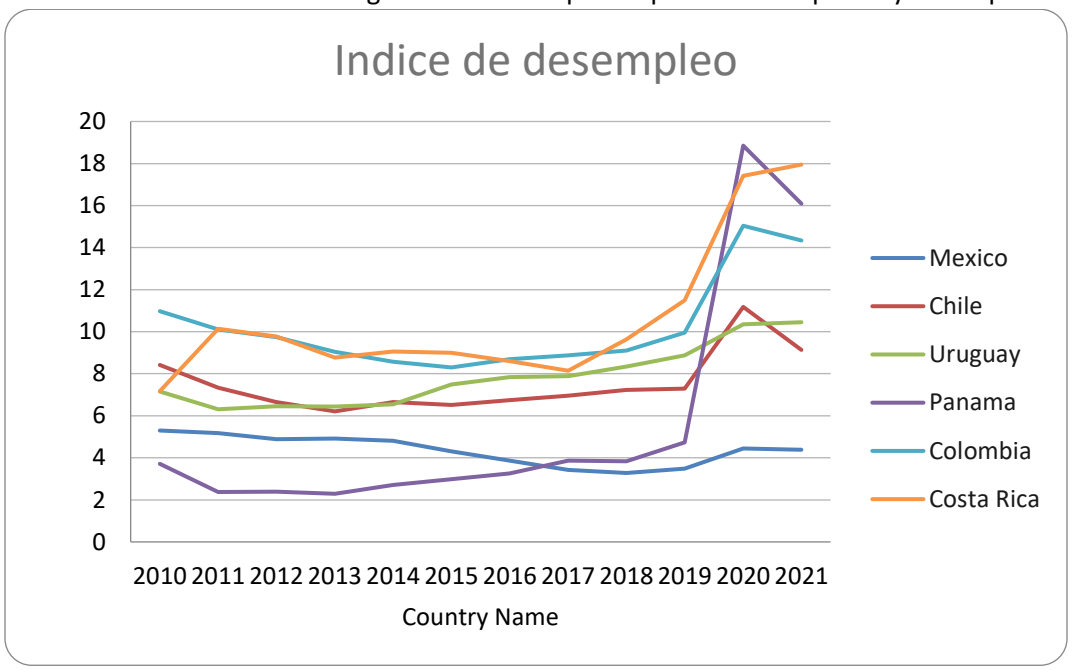
4. Índice de desempleo:
   Descripción de la aproximación: Curiosidad
   Descripción de las variables:

   - Nombre del país: Texto
     Esta variable representa el nombre del país al que pertenecen los datos registrados en cada fila. Cada fila probablemente corresponde a un país específico, y esta columna proporciona el nombre de ese país.

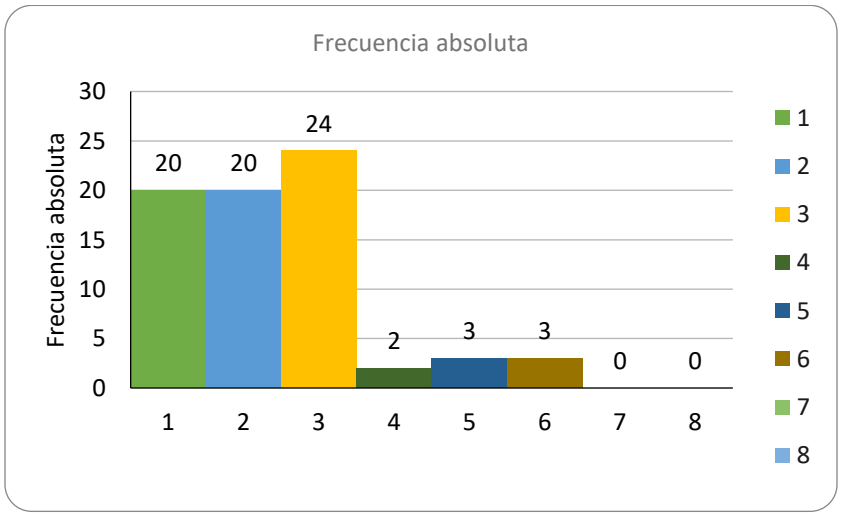
   - Código del país: Numérico
     Esta variable contiene el código asignado al país correspondiente en la columna "Nombre del país". Los códigos de países son identificadores numéricos o alfanuméricos utilizados para identificar y diferenciar países de manera única en distintos sistemas y bases de datos.

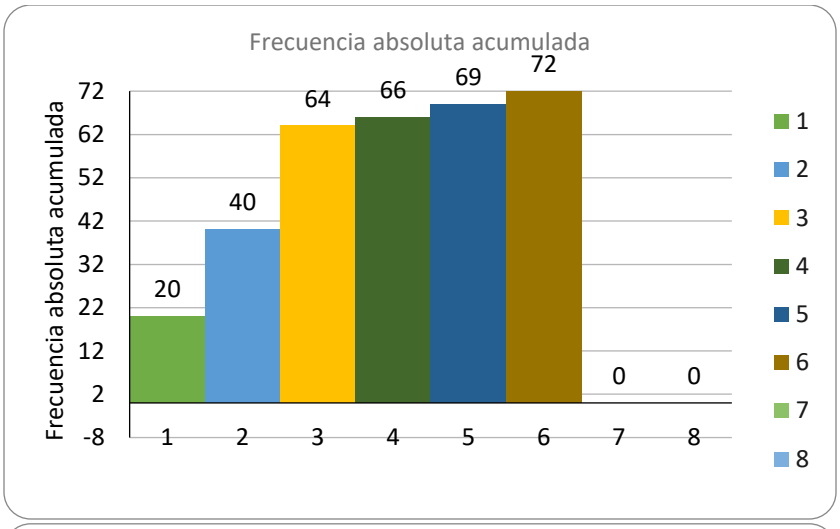
   - 2010, 2011, 2012, ..., 2021: Numérico
     Estas columnas representan los datos para cada año desde 2010 hasta 2021. Cada columna contiene información específica que probablemente se refiere a indicadores, estadísticas o valores relacionados con el país en cuestión. El tipo de datos en estas columnas dependerá del contexto de la base de datos.

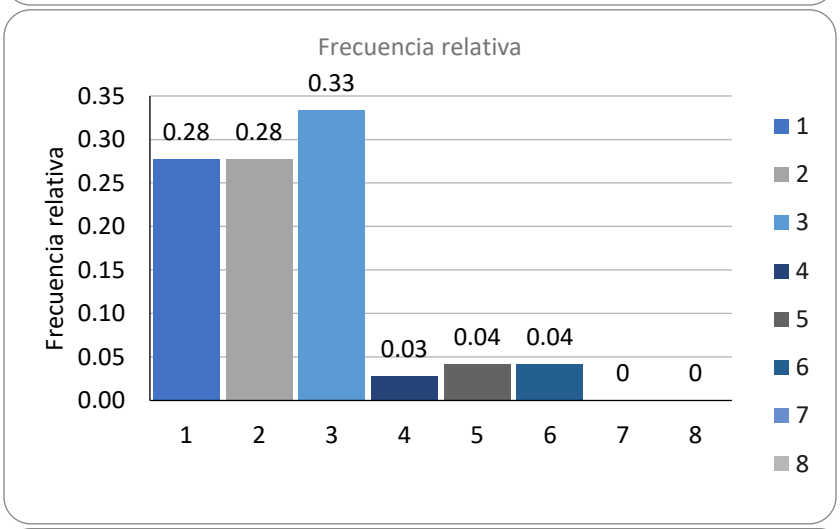
Podemos observar un gráfico de líneas para representar los países y los respectivos años

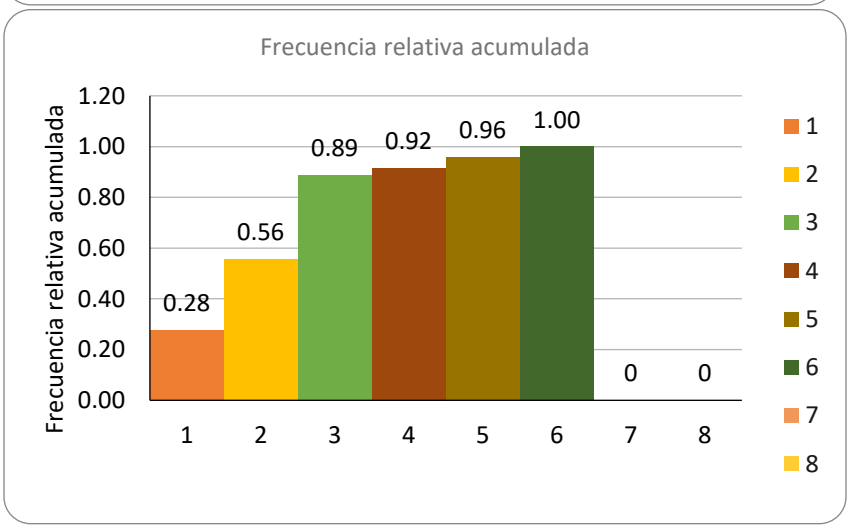


Intervalo	Frecuencia absoluta	Frecuencia absoluta acumulada	Frecuencia relativa	Frecuencia relativa acumulada	Media	varianza	desviacion estandar
2-4	20	20	0.28	0.28	**7.69**	**13.41**	**3.66143946**
5-7	20	40	0.28	0.56			
8-10	24	64	0.33	0.89			
11-13	2	66	0.03	0.92			
14-16	3	69	0.04	0.96			
17-19	3	72	0.04	1.00			




Frecuencia absoluta acumulada
Frecuencia absoluta acumulada
72
62
52
42
32
22
12
2
-8
20
40
64
66
69
72
0
0
1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8


Frecuencia relativa
Frecuencia relativa
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.28
0.28
0.33
0.03
0.04
0.04
0
0
1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8


Frecuencia relativa acumulada
Frecuencia relativa acumulada
1.20
1.00
0.80
0.60
0.40
0.20
0.00
0.28
0.56
0.89
0.92
0.96
1.00
0
0
1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8

Tendencia del Índice de Desempleo:

Regression Statistics	
Multiple R	0.699176374
R Square	0.488847602
Adjusted R Square	0.437732362
Standard Error	4.219860807
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	170.3018393	170.3018393	9.563637069	0.011395071
Residual	10	178.0722523	17.80722523		
Total	11	348.3740917			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
Intercept	-2193.911632	711.2356985	-3.084647799	0.01154882	-3778.643518	-609.1797458	-3778.643518	-609.1797458
X Variable 1	1.091293706	0.352882488	3.092513067	0.011395071	0.305022527	1.877564885	0.305022527	1.877564885

A lo largo de los años, el índice de desempleo ha mostrado una variación significativa. Iniciando en el año 2010 con una tasa de desempleo del 3.72%, ha habido fluctuaciones, alcanzando su punto más bajo en el año 2013 con un 2.29%. Sin embargo, desde el año 2014, la tendencia ha sido al alza, alcanzando su máximo en el año 2020 con un 18.85% y disminuyendo ligeramente en el año 2021 con un 16.09%.

Correlación y Ajuste del Modelo:

El coeficiente de correlación (Multiple R) entre el año y el índice de desempleo es de aproximadamente 0.699. Esto sugiere una correlación moderada entre ambas variables, lo que indica que el año puede tener cierta influencia en las tasas de desempleo.

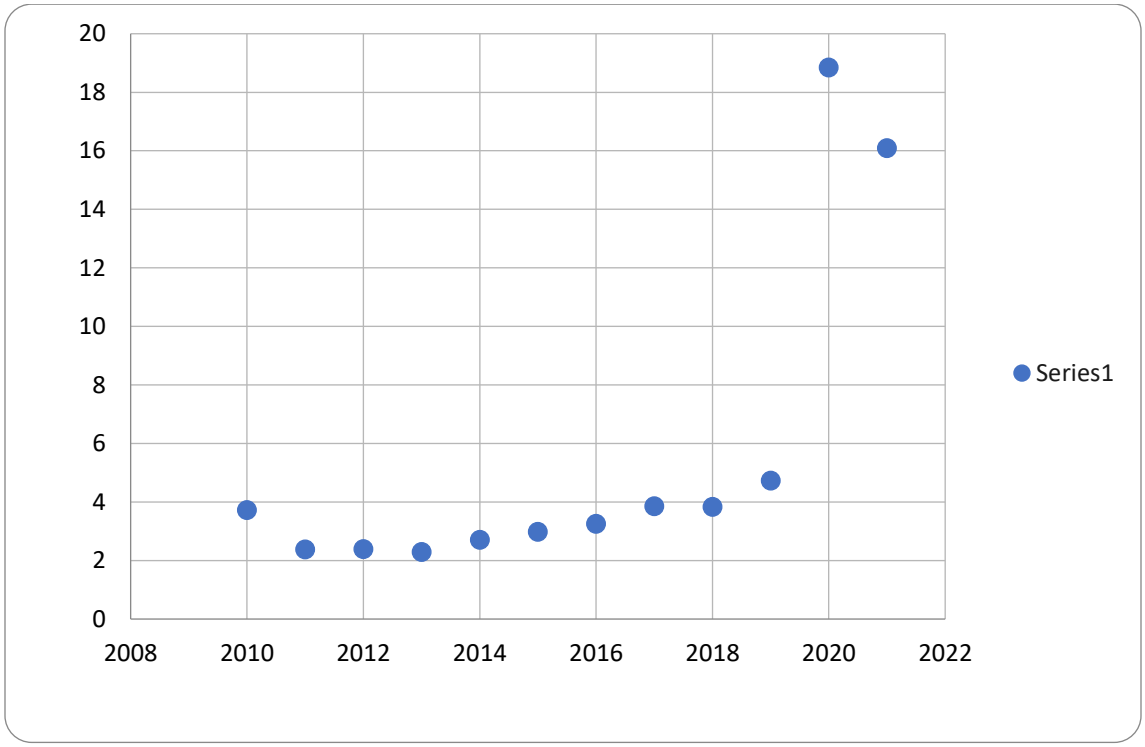
El coeficiente de determinación (R cuadrado) es de aproximadamente 0.489, lo que significa que aproximadamente el 48.9% de la variabilidad en el índice de desempleo puede ser explicada por el año. Esto indica que el modelo de regresión puede explicar una parte significativa de la variabilidad en las tasas de desempleo.

Error Estándar:

El error estándar es de aproximadamente 4.22. Esto indica que la diferencia entre los valores reales y los valores predichos por el modelo de regresión podría ser de alrededor de 4.22 puntos porcentuales. Un error estándar bajo indica una mejor precisión del modelo.

	año	*Indice desempleo*
año	1	
Indice desempleo	0.699176374	1

La correlación entre las variables "Año" e "Índice Desempleo" es de aproximadamente 0.699, lo que indica una correlación positiva moderada entre ambos. Esto sugiere que, en general, a lo largo de los años, ha habido una tendencia al aumento del índice de desempleo. Sin embargo, es importante mencionar que, en los años 2020 y 2021, se observan aumentos drásticos en el índice de desempleo en comparación con los años anteriores, lo que puede indicar un impacto significativo causado por eventos específicos, como la pandemia del COVID-19.



5. Valores PIB interno bruto Anual
   Descripción de la aproximación: Curiosidad
   Descripcion de las variables:
   - Prefijo:
     Esta variable puede ser un código o un prefijo que se utiliza para identificar o agrupar ciertas categorías o datos relacionados en la base de datos. Sirve como un identificador inicial para ciertos registros.
   - Categorías:
     Esta variable representa las diferentes categorías o clasificaciones de datos en la base de datos. Puede ser cualquier tipo de agrupación, como tipos de productos, industrias, sectores, regiones, etc.
   - Años/Relación - Años:
     Estas variables se refieren a años específicos o períodos de tiempo y pueden estar relacionadas entre sí de alguna manera. Pueden utilizarse para mostrar relaciones temporales o comparaciones entre diferentes años.
   - Año:
     Esta variable representa un año específico en la línea de tiempo y suele utilizarse como etiqueta temporal para cada conjunto de datos correspondiente a un año específico.
   - Codcategoria:
     Esta variable puede ser un código numérico o alfanumérico que identifica a cada categoría de manera única. Se utiliza como un identificador para cada categoría en la base de datos.
   - Relación - Año:
     Esta variable podría representar una relación o comparación con un año específico, o tal vez muestra una proporción o tasa relacionada con ese año.
   - Composición Constante:
     Esta variable puede referirse a una composición o estructura específica de datos que se mantiene constante durante un período de tiempo determinado.
   - Composición Corriente:
     Esta variable puede referirse a una composición o estructura específica de datos en el presente o en el año actual.
   - Valor Constante:
     Esta variable representa un valor numérico constante o invariable a lo largo de un período de tiempo específico.
   - Valor Corriente:
     Esta variable representa un valor numérico actual o correspondiente al año actual.
   - Valores:
     Esta variable puede contener diferentes valores numéricos relacionados con las categorías o los años.

Variación Absoluta Constante:

Esta variable muestra la diferencia o cambio absoluto entre dos valores constantes o invariables.

Variación Absoluta Corriente:

Esta variable muestra la diferencia o cambio absoluto entre dos valores correspondientes al año actual.

Variación Porcentual Constante:

Esta variable representa la variación o cambio porcentual entre dos valores constantes o invariables.

Variación Porcentual Corriente:

Esta variable representa la variación o cambio porcentual entre dos valores correspondientes al año actual.

Intervalos de años	Total	Frecuencia	Frecuencia absoluta	Frecuencia abs. Acumulativa	Frecuencia Relativa	Frecuencia Relativa Acumulativa
2007-2009	15858	5	5	5	16.13	16.13
2010-2011	45619	11	11	16	35.48	51.61
2012-2013	5449	2	2	18	6.45	58.06
2014-2015	79710	5	5	23	16.13	74.19
2016-2017	75222	6	6	29	19.35	93.55
2018-2019	60629	0	0	29	0.00	93.55
2020	3093	2	2	31	6.45	100
Total	285580	31	31	151	100	487.096774
Media	40797.1429	4.42857143	4.42857143	21.5714286	14.2857143	69.5852535

Distribución de datos en intervalos de años:

Los datos se han agrupado en diferentes intervalos de años, desde 2007 hasta 2020. Cada intervalo representa un rango de años específico.

Frecuencia absoluta:

La frecuencia absoluta representa el número de observaciones o datos que se encuentran dentro de cada intervalo. Por ejemplo, el intervalo 2007-2009 tiene 5 observaciones, el intervalo 2010-2011 tiene 11 observaciones, etc.

Frecuencia absoluta acumulativa:

La frecuencia absoluta acumulativa muestra la suma acumulativa de las frecuencias absolutas a medida que avanzamos a través de los intervalos. Por ejemplo, después de los intervalos 2014-2015, la frecuencia absoluta acumulativa es de 23, lo que indica que en total hay 23 observaciones hasta ese punto.

Frecuencia relativa:

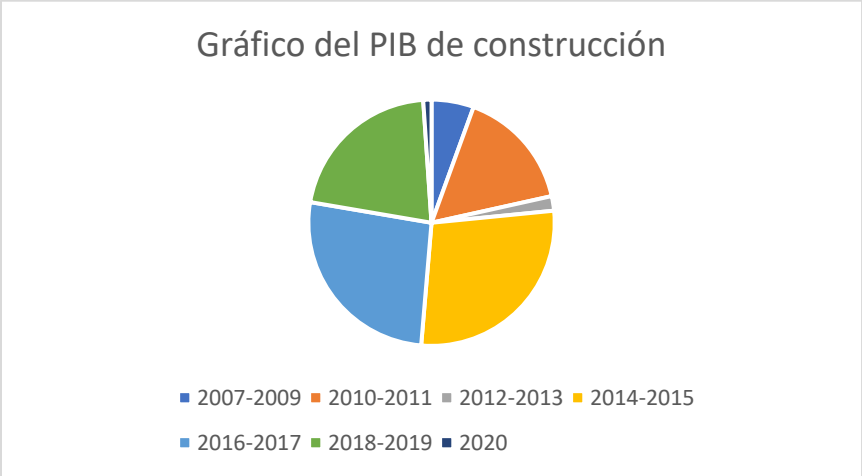
La frecuencia relativa representa el porcentaje de observaciones en cada intervalo con respecto al total de observaciones. Por ejemplo, el intervalo 2007-2009 tiene una frecuencia relativa del 16.13%, lo que significa que representa aproximadamente el 16.13% del total de observaciones.

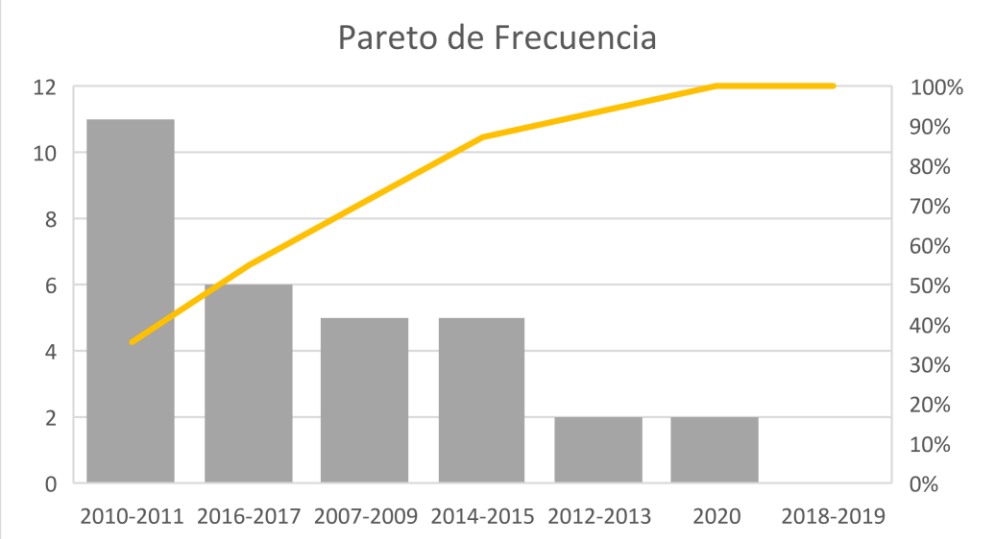
Frecuencia relativa acumulativa:

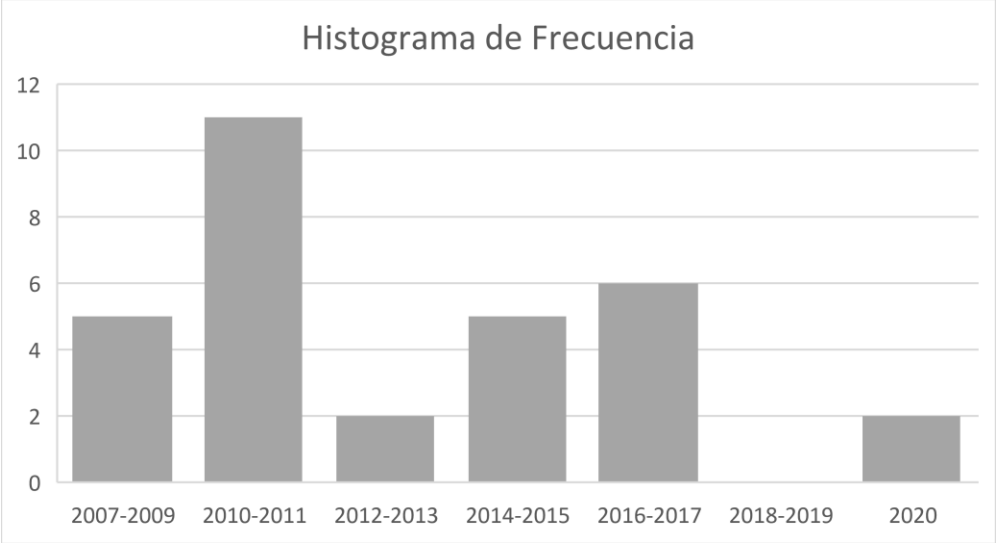
La frecuencia relativa acumulativa muestra la suma acumulativa de las frecuencias relativas a medida que avanzamos a través de los intervalos. Por ejemplo, después de los intervalos 2016-2017, la frecuencia relativa acumulativa es de 93.55%, lo que indica que hemos cubierto aproximadamente el 93.55% de las observaciones hasta ese punto.

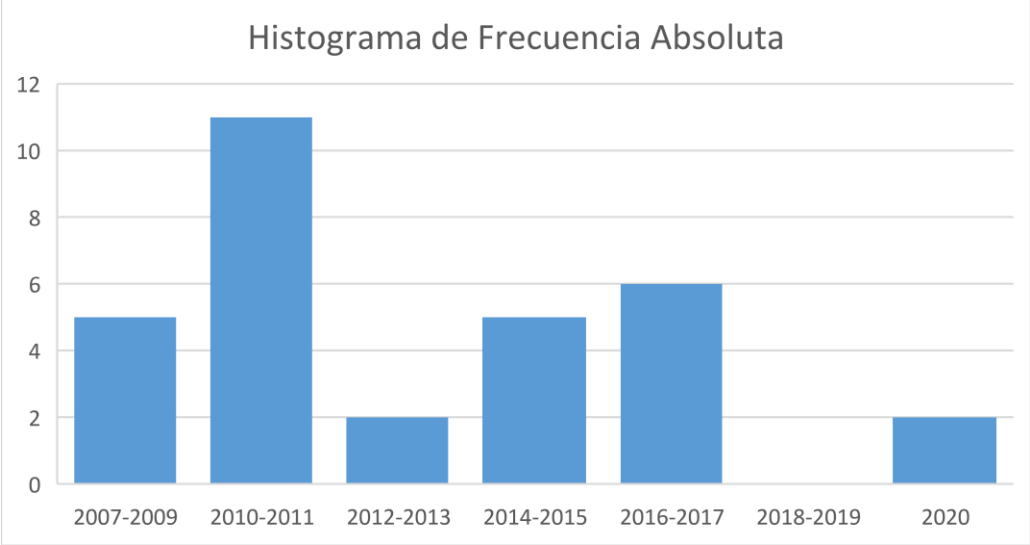
Media:

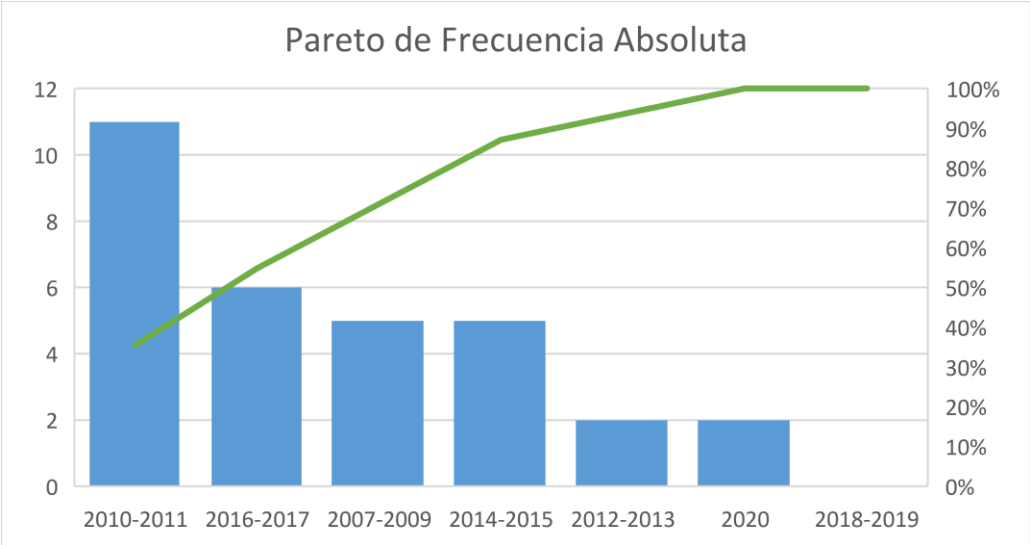
La media representa el valor promedio de los datos. En este caso, la media de los intervalos de años es de aproximadamente 40,797.14.

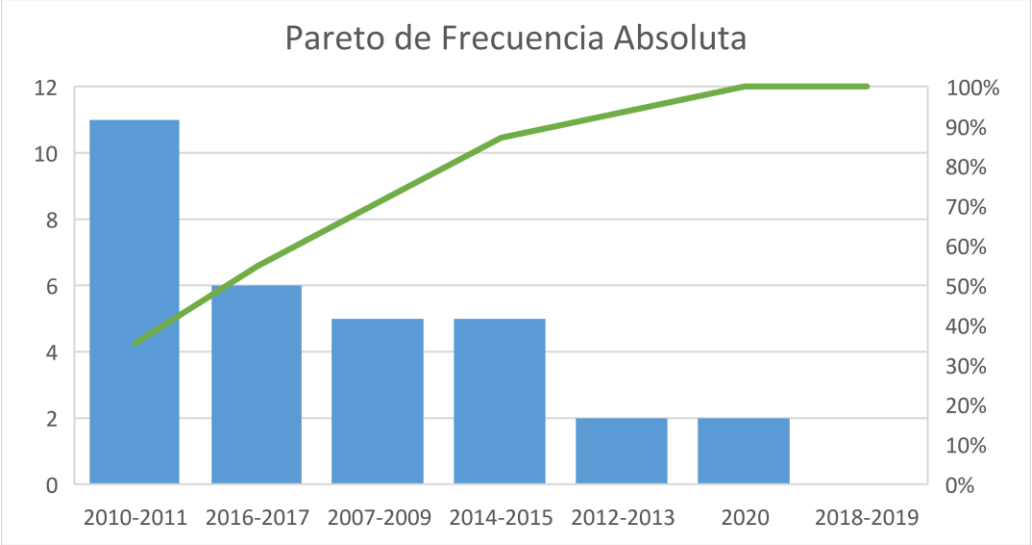


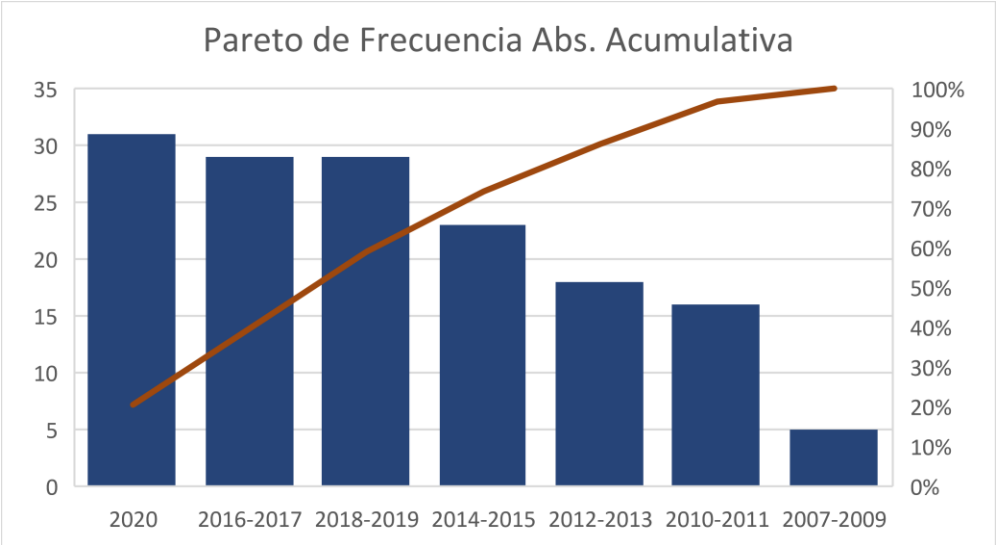


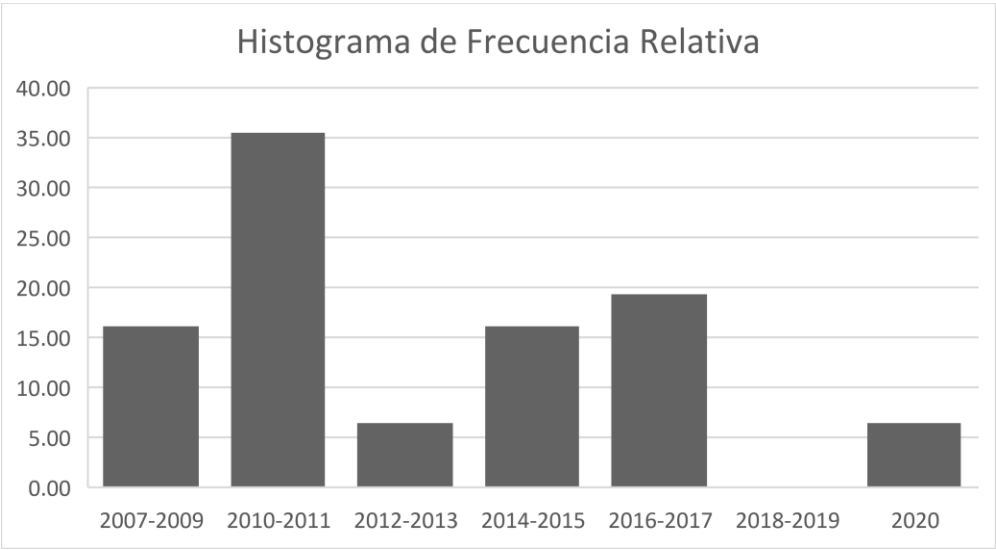

Histograma de Frecuencia
12
10
8
6
4
2
0
2007-2009
2010-2011
2012-2013
2014-2015
2016-2017
2018-2019
2020


Histograma de Frecuencia Absoluta
12
10
8
6
4
2
0
2007-2009
2010-2011
2012-2013
2014-2015
2016-2017
2018-2019
2020


Pareto de Frecuencia Absoluta
12
10
8
6
4
2
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2010-2011
2016-2017
2007-2009
2014-2015
2012-2013
2020
2018-2019


Pareto de Frecuencia Absoluta
12
10
8
6
4
2
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2010-2011
2016-2017
2007-2009
2014-2015
2012-2013
2020
2018-2019


Pareto de Frecuencia Abs. Acumulativa
35
30
25
20
15
10
5
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2020
2016-2017
2018-2019
2014-2015
2012-2013
2010-2011
2007-2009


Histograma de Frecuencia Relativa
40.00
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
2007-2009
2010-2011
2012-2013
2014-2015
2016-2017
2018-2019
2020

```python
import numpy as np

# Datos discretos de construcción
datos_discretos = [15858, 45619, 5449, 79710, 75222, 60629, 3093]

# Transformación logarítmica
datos_transformados_log = np.log(datos_discretos)

# Transformación raíz cuadrada
datos_transformados_sqrt = np.sqrt(datos_discretos)

print("Datos discretos de construcción:", datos_discretos)
print("Transformación logarítmica:", datos_transformados_log)
print("Transformación raíz cuadrada:", datos_transformados_sqrt)
```

6. Indice de felicidad
   Descripción de la aproximación: Curiosidad
   Descripción de variables:

   Clasificación general:
   Esta variable representa la posición o el rango en el que se encuentra cada país o región en una clasificación general o ranking. La clasificación general podría estar relacionada con un índice o medida específica que evalúa diferentes aspectos o indicadores de cada país o región.

   País o Región:
   Esta variable indica el nombre del país o la región a la que pertenecen los datos registrados en cada fila. Cada fila probablemente corresponde a un país o región específica, y esta columna proporciona el nombre de ese país o región.

   Año:
   Esta variable representa el año en el que se registraron los datos. Es un dato temporal que proporciona información sobre el período al que pertenece cada conjunto de datos.

   Puntaje:
   Esta variable contiene un valor numérico que representa un puntaje o calificación asociada al país o región en la clasificación general. El puntaje podría ser el resultado de un índice compuesto o una medida específica que evalúa el rendimiento en diferentes aspectos.

   PIB per cápita:
   Esta variable representa el Producto Interno Bruto (PIB) per cápita del país o región. El PIB per cápita es una medida económica que indica el valor promedio del producto

interno bruto dividido por la población total. Es una indicación del nivel de riqueza o bienestar económico de un país.

Apoyo social:
Esta variable puede representar el nivel de apoyo social percibido o experimentado por los ciudadanos en el país o región. Puede estar relacionada con aspectos como el sistema de seguridad social, programas de bienestar, redes de apoyo comunitario, entre otros.

Esperanza de vida saludable: Esta variable indica la esperanza de vida saludable en el país o región. Representa el número promedio de años que se espera que una persona viva en buen estado de salud, sin enfermedades graves o discapacidades.

Libertad para tomar decisiones en la vida:
Esta variable puede medir el grado de libertad que los ciudadanos tienen para tomar decisiones en sus vidas diarias. Puede estar relacionada con aspectos como la libertad de expresión, de asociación, de movimiento, entre otros.

Generosidad:
Esta variable puede representar la generosidad percibida o medida en el país o región. Puede estar relacionada con actos de caridad, donaciones, o el grado de disposición a ayudar a otros.

Percepciones de corrupción:
Esta variable puede medir las percepciones o la percepción pública sobre el nivel de corrupción en el país o región. Representa cómo los ciudadanos perciben la existencia y prevalencia de la corrupción en diferentes ámbitos de la sociedad.

Intervalo	Frecuencia absoluta	Fr. absoluta acumulada	Frecuencia relativa	Fr. relativa acumulada	**Media**	**Desviación estandar**	**Varianza**
3,1 - 3.5	17	17	0,06	0,06	209,41	1,10	1,22
3.6 - 4	15	32	0,05	0,10			
4,1 - 4.5	52	84	0,17	0,27			
5,1 - 5.5	50	169	0,16	0,55			
5.,6 - 6	50	219	0,16	0,72			
6,1 - 6.5	44	263	0,14	0,86			
6.6 - 7	19	282	0,06	0,92			
7,1 - 7,5	22	304	0,07	0,99			
7,6 - 8	2	306	0,01	1,00			

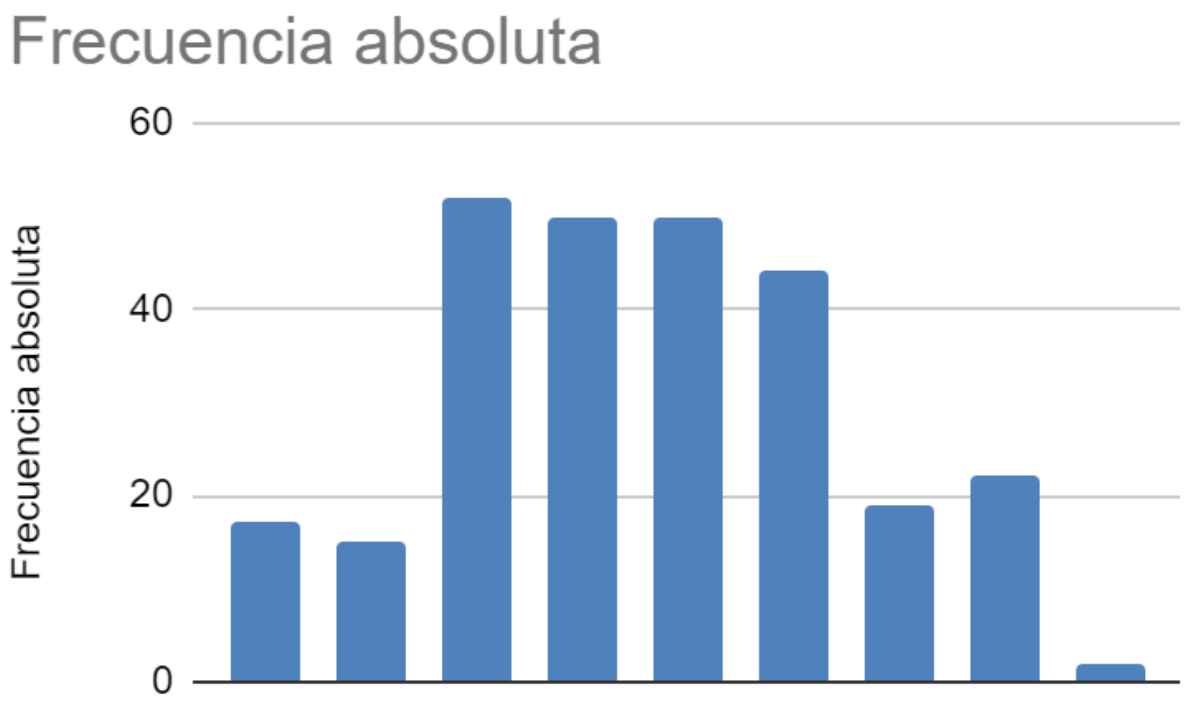
La mayoría de las observaciones del índice de felicidad se encuentran en los intervalos entre 4.1 y 5.5, lo que indica que la mayoría de las personas tienen un índice de felicidad en ese rango.

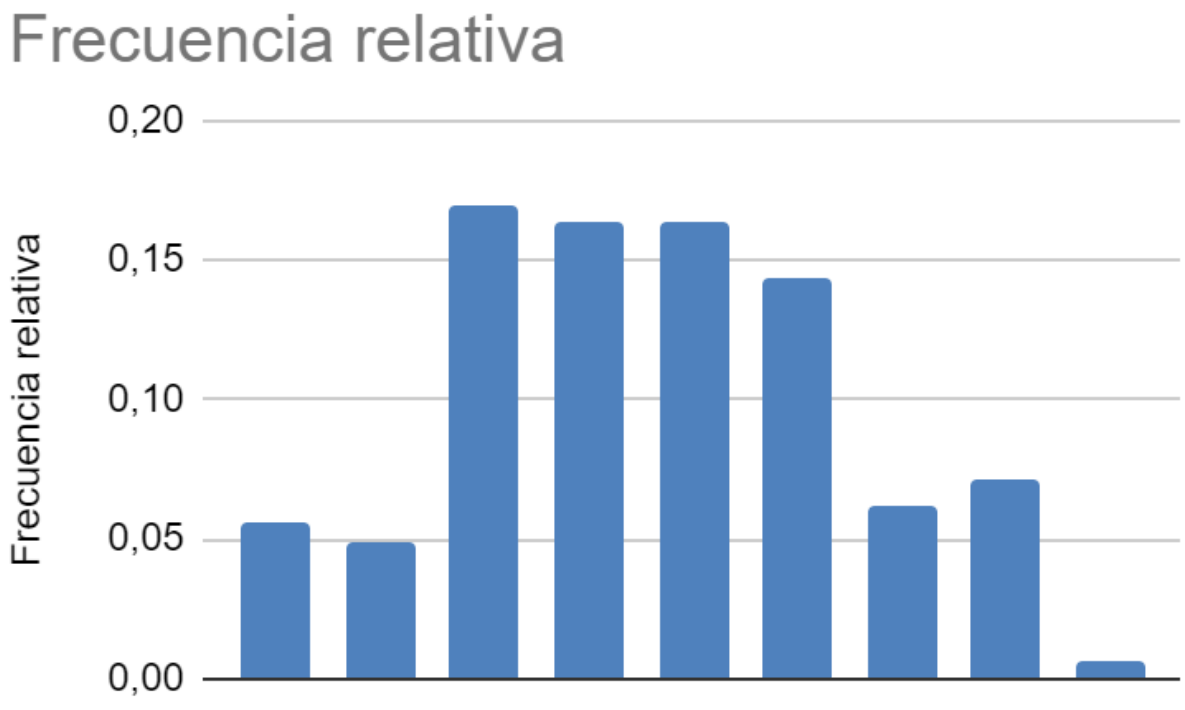
El promedio (media) del índice de felicidad es de aproximadamente 209.41.

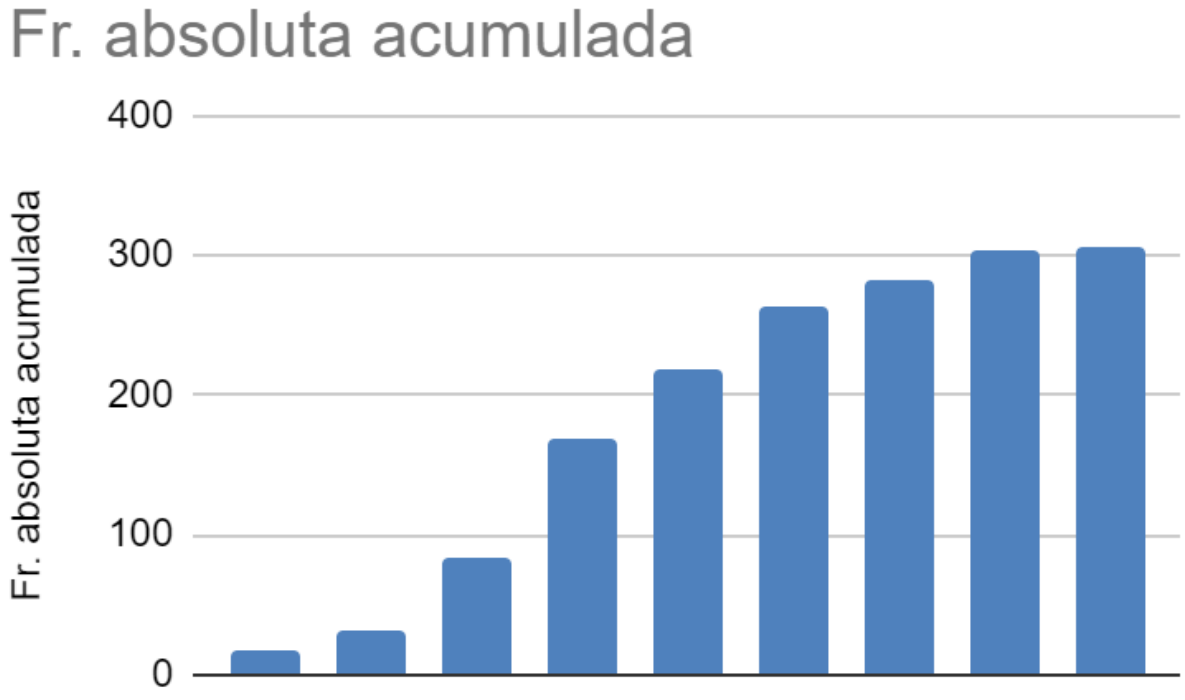
La desviación estándar del índice de felicidad es de aproximadamente 1.10, lo que sugiere que los valores tienden a estar cerca del promedio.

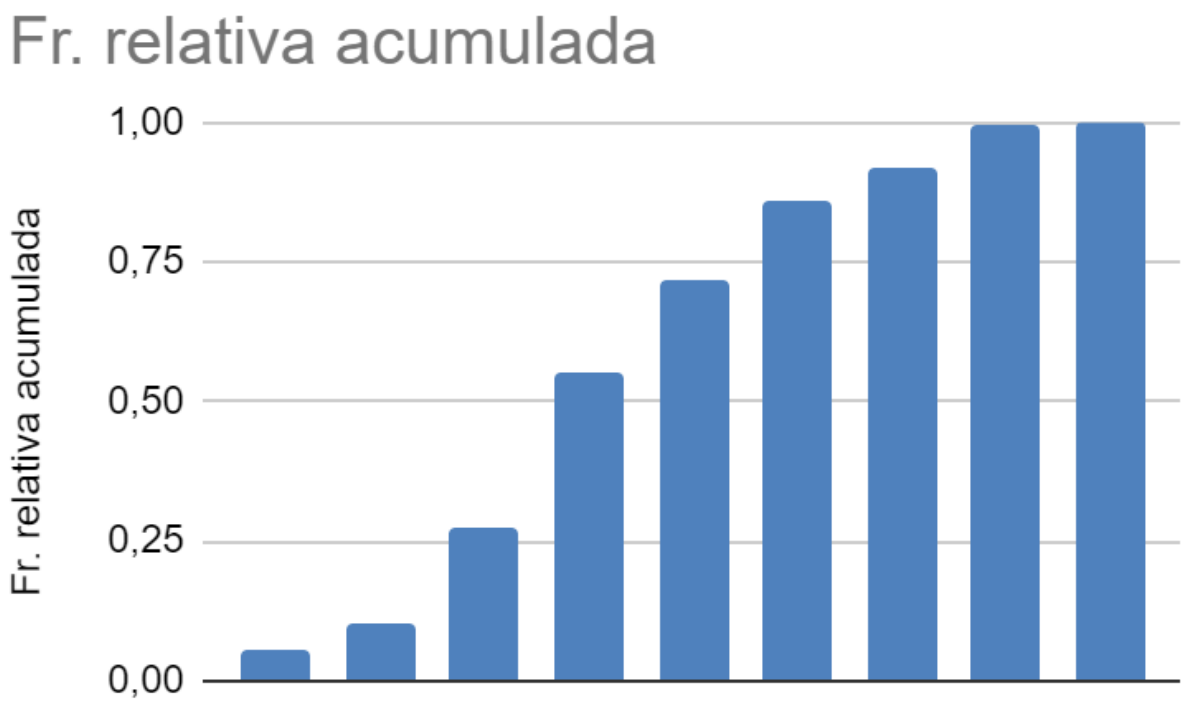
La varianza del índice de felicidad es de aproximadamente 1.22, lo que indica una pequeña dispersión en los valores con respecto a la media.

La frecuencia relativa acumulada alcanza el 100%, lo que significa que hemos abarcado todas las observaciones de los datos en los intervalos.




Frecuencia relativa
0,20
0,15
0,10
0,05
0,00
Frecuencia relativa


Fr. absoluta acumulada
400
300
200
100
0
Fr. absoluta acumulada



Regression Statistics	
Multiple R	0,9845043776
R Square	0,9692488694
Adjusted R Square	0,969054242
Standard Error	0,1982696397
Observations	160

ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	195,7689167	195,7689167	4980,022475	0			
Residual	158	6,211114305	0,03931085003					
Total	159	201,980031						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
Intercept	0,03169274 25	0,07734533179	0,4097563 715	0,6825396 981	-0,1210714 064	0,1844568 914	-0,1210714 064	0,1844568 914
X Variable 1	0,9876064 745	0,0139948 5094	70,569274 3	0	0,9599653 558	1,0152475 93	0,9599653 558	1,0152475 93

Coeficiente de correlación (Multiple R):

El coeficiente de correlación múltiple, que es 0.9845, indica una fuerte correlación positiva entre las variables analizadas en el modelo de regresión. Esto sugiere que existe una relación lineal significativa entre las variables.

Coeficiente de determinación (R Square):

El coeficiente de determinación, que es 0.9692, indica que aproximadamente el 96.92% de la variabilidad en la variable dependiente puede explicarse por la variable independiente en el modelo de regresión. Es decir, el modelo de regresión es capaz de explicar gran parte de la variabilidad observada en los datos.

Coeficiente de determinación ajustado (Adjusted R Square):

El coeficiente de determinación ajustado, que es 0.9691, es una versión corregida del R Square que tiene en cuenta la cantidad de variables independientes en el modelo. Indica que el modelo ajustado también es capaz de explicar una gran cantidad de la variabilidad en los datos.

Error estándar (Standard Error):

El error estándar, que es 0.1983, representa la estimación de la desviación estándar de los residuos en el modelo de regresión. Cuanto menor sea el error estándar, mejor se ajustará el modelo a los datos.

Tamaño de la muestra (Observations):

La muestra contiene 160 observaciones.

Prueba de ANOVA:

La prueba de ANOVA muestra que el modelo de regresión es significativo con un valor de p (Significance F) igual a 0. Esto indica que al menos una de las variables independientes es significativa para explicar la variabilidad de la variable dependiente.

Coeficientes del modelo:

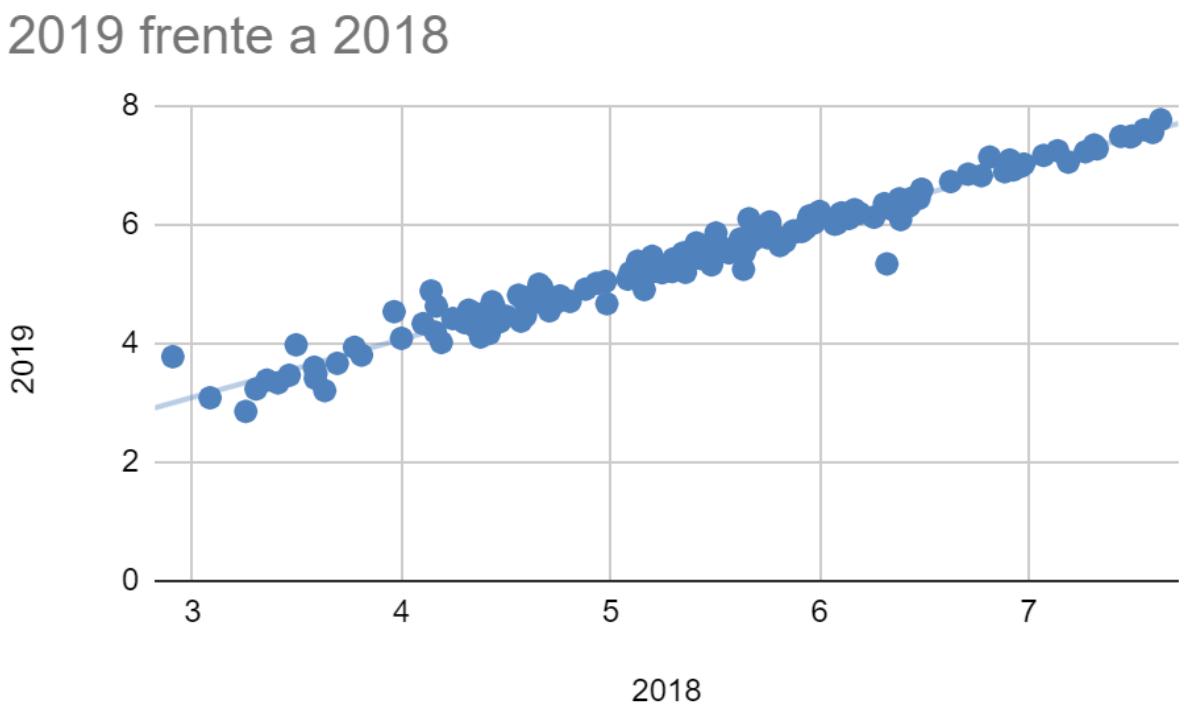
El intercepto (0.0317) y el coeficiente de la variable X (0.9876) indican el modelo de regresión lineal:

$Y = 0.0317 + 0.9876 * X$

El coeficiente de la variable X (X Variable 1) sugiere que, por cada incremento de una unidad en X, la variable dependiente Y aumenta en aproximadamente 0.9876 unidades.

Análisis de Correlación

	2018	*2019*
2018	1	
2019	0,9838422458	1



La correlación de 0.9838 en el año 2019 sugiere que existe una relación muy estrecha y positiva entre las dos variables analizadas. Una correlación cercana a 1 indica que las variables tienden a moverse en la misma dirección, es decir, cuando una variable aumenta, la otra también tiende a aumentar, y cuando una variable disminuye, la otra también tiende a disminuir.

7. Base de Datos:
   Tipo de Análisis: Curiosidad
   EDIFICACIONES, ÁREA Y COSTO DE LAS CONSTRUCCIONES, ADICIONES Y REPARACIONES PARTICULARES EN ALGUNOS DISTRITOS DE LA REPÚBLICA, POR CLASE: ENERO A ABRIL 2022-23
   Distrito:

   Esta variable indica el nombre o identificador del distrito específico al que pertenecen los datos registrados en cada fila. Cada fila probablemente corresponde a un distrito diferente, y esta columna proporciona el nombre o identificación de ese distrito.
   Edificaciones - Total:

   Representa el número total de edificaciones registradas en el distrito. Las edificaciones incluyen tanto edificios residenciales como no residenciales.

Área (En m2) - Total:

Indica el área total, en metros cuadrados (m2), de todas las edificaciones registradas en el distrito, incluyendo tanto las edificaciones residenciales como las no residenciales.
Costo (En balboas) - Total (1):

Esta variable muestra el costo total, en balboas (moneda de Panamá), de todas las edificaciones registradas en el distrito, incluyendo tanto las edificaciones residenciales como las no residenciales.
Edificaciones - Residenciales:

Representa el número total de edificaciones residenciales registradas en el distrito.
Unidades de vivienda (3):

Esta variable indica el número total de unidades de vivienda registradas en el distrito. Las unidades de vivienda se refieren a las unidades habitables dentro de las edificaciones residenciales.
Área (En m2) - Residenciales:

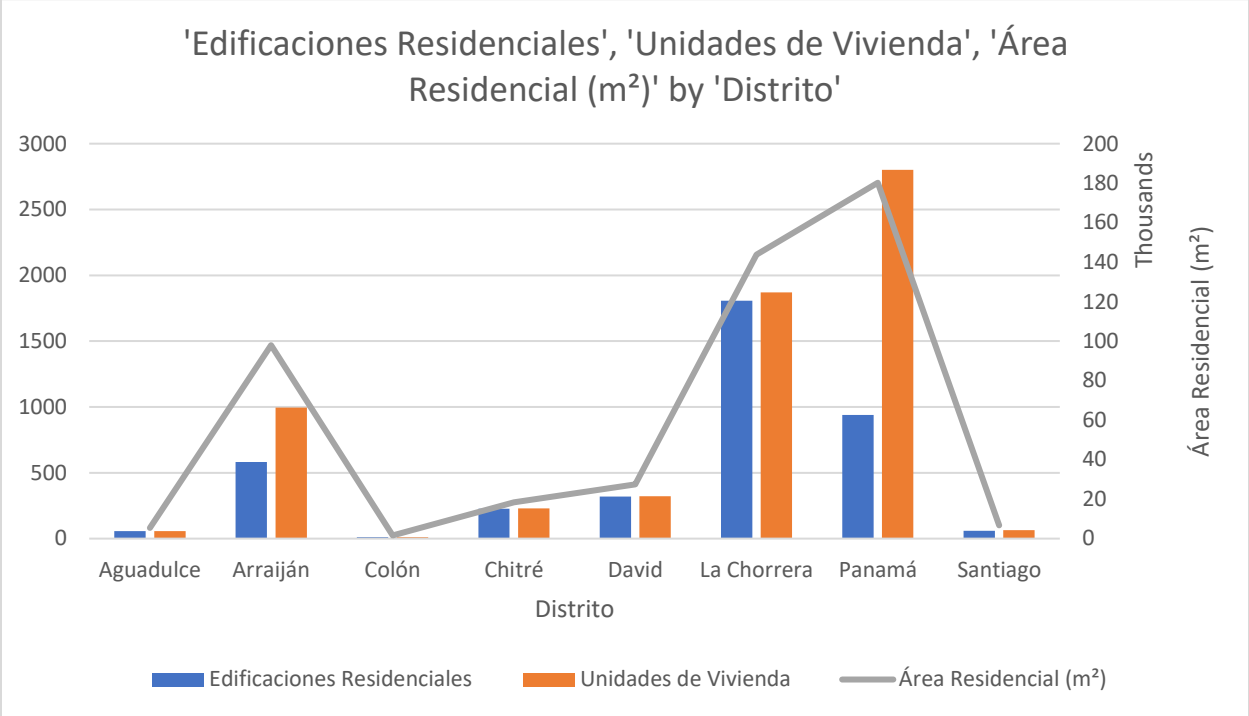
Indica el área total, en metros cuadrados (m2), de las edificaciones residenciales registradas en el distrito.
Costo (En balboas) - Residenciales (1):

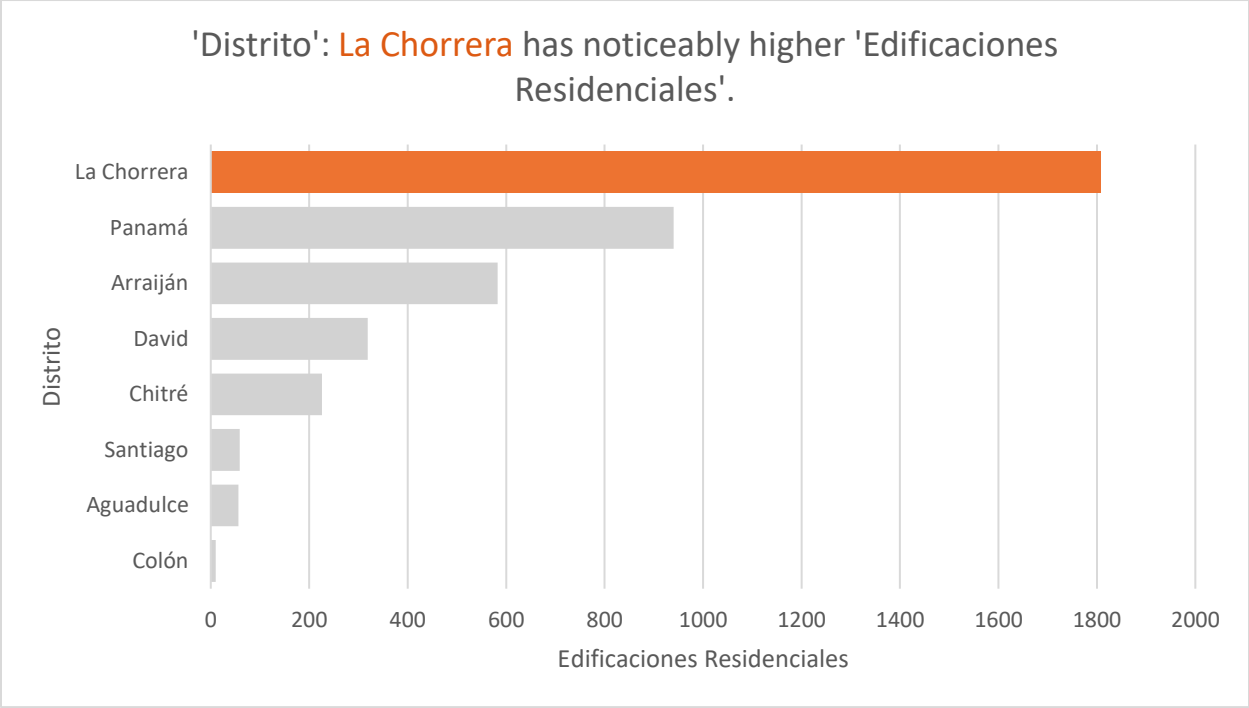
Esta variable muestra el costo total, en balboas (moneda de Panamá), de todas las edificaciones residenciales registradas en el distrito.
Edificaciones - No residenciales:

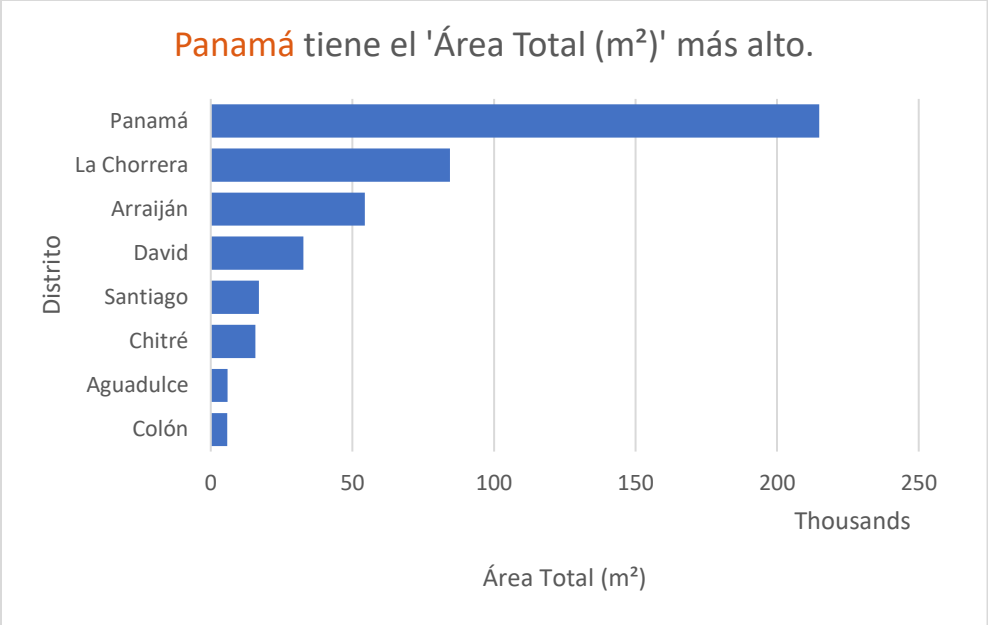
Representa el número total de edificaciones no residenciales registradas en el distrito.
Área (En m2) - No residenciales:

Indica el área total, en metros cuadrados (m2), de las edificaciones no residenciales registradas en el distrito.
Costo (En balboas) - No residenciales (1):

Esta variable muestra el costo total, en balboas (moneda de Panamá), de todas las edificaciones no residenciales registradas en el distrito.


'Edificaciones Residenciales', 'Unidades de Vivienda', 'Área Residencial (m²)' by 'Distrito'
3000
2500
2000
1500
1000
500
0
200
180
160
140
120
100
80
60
40
20
0
Thousands
Área Residencial (m²)
Aguadulce
Arraiján
Colón
Chitré
David
La Chorrera
Panamá
Santiago
Distrito
Edificaciones Residenciales
Unidades de Vivienda
Área Residencial (m²)


'Distrito': La Chorrera has noticeably higher 'Edificaciones Residenciales'.
La Chorrera
Panamá
Arraiján
David
Chitré
Santiago
Aguadulce
Colón
Distrito
0
200
400
600
800
1000
1200
1400
1600
1800
2000
Edificaciones Residenciales



Conclusiones:

Distribución de edificaciones: Los distritos de La Chorrera, Panamá y David tienen la mayor cantidad de edificaciones, con 852, 1,049 y 351 respectivamente. Mientras que los distritos de Colón y Aguadulce tienen la menor cantidad, con solo 17 y 65 edificaciones respectivamente.
Medidas de tendencia central: La media de edificaciones totales es de aproximadamente 376. Esto indica que, en promedio, los distritos tienen alrededor de 376 edificaciones. La moda es 0, lo que sugiere que hay al menos un distrito sin edificaciones. La mediana es 245.5, lo que indica que la mitad de los distritos tienen menos de 245.5 edificaciones y la otra mitad tiene más.
Rango y dispersión: El rango de edificaciones totales es de 1032, lo que significa que hay una diferencia significativa entre el distrito con la mayor cantidad de edificaciones y el distrito con la menor cantidad. La varianza es de 147022.5 y la desviación estándar es de 383.4351314, lo que indica una dispersión considerable de los datos alrededor de la media.

Recomendaciones:

Análisis de outliers: Dado que la moda es 0 y el rango es considerable, se recomienda investigar y analizar el distrito con 0 edificaciones para comprender las razones detrás de este valor atípico.
Análisis de los distritos más y menos desarrollados: Puedes realizar un análisis más detallado de los distritos de La Chorrera, Panamá y David, que tienen la mayor cantidad de edificaciones, para identificar patrones o características comunes que contribuyan a su mayor desarrollo. Del mismo modo, puedes analizar los distritos de Colón y Aguadulce para comprender las razones detrás de su menor cantidad de edificaciones.
Considerar factores adicionales: Además del número de edificaciones, es importante considerar otros factores relevantes, como el tamaño de las edificaciones, el uso (residencial o no residencial), la infraestructura disponible y la demanda en cada distrito. Estos factores pueden ayudar a comprender mejor las diferencias observadas en la cantidad de edificaciones y proporcionar una visión más completa del panorama de construcción en cada distrito.

Distrito	Edificaciones totales	Media	Moda	Mediana	Rango	Varianza	Desviacion Estandar
Aguadulce	65	376	0	245.5	1032	147022.5	383.4351314
Arraiján	417						
Colón	17						
Chitré	119						
David	351						
La Chorrera	852						
Panamá	1,049						
Santiago	140						

Regression Statistics	
Multiple R	0.914922415
R Square	0.837083026
Adjusted R Square	0.804499631
Standard Error	172.9911548
Observations	7

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	768811.7	768811.7	25.69048	0.003872
Residual	5	149629.7	29925.94		
Total	6	918441.4			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*U 9*
Intercept	123.5107123	87.8258	1.406315	0.218633	-102.253	349.2741	-102.253	3
5885	0.004894484	0.000966	5.068578	0.003872	0.002412	0.007377	0.002412	0

El modelo de regresión tiene un buen ajuste y explica una gran parte de la variabilidad en la variable dependiente, con un R Square de 0.8371.

La variable independiente "5885" parece tener un efecto significativo en la variable dependiente, como se indica por el valor de p (P-value) en la prueba de ANOVA y el coeficiente asociado en los resultados del modelo.

Sin embargo, es importante tener en cuenta que estas conclusiones se basan en los resultados presentados y que un análisis más detallado requeriría una revisión completa del contexto, el significado de las variables y una evaluación adicional de la bondad del ajuste del modelo.

8. Base de datos:
   Descripción de la aproximación: Curiosidad
   Exportaciones de los principales países
   - Descripcion de variables:
     Dentro/fuera de País conjunto:

       Esta variable indica si los datos corresponden a transacciones comerciales realizadas dentro del país (dentro de las fronteras del país) o fuera del país (transacciones internacionales o de comercio exterior).
   - Grupo:

       Esta variable puede representar una categoría o clasificación específica a la que pertenecen los datos. Por ejemplo, podría ser una clasificación de productos, sectores industriales o cualquier otra agrupación relevante para el análisis.
   - Países:

       Esta variable hace referencia a los países involucrados en las transacciones comerciales. Puede haber un país de origen (si se trata de importaciones) y un país de destino (si se trata de exportaciones) dependiendo del contexto de los datos.
   - Años:

       Esta variable indica el año en el que se realizaron las transacciones comerciales. Representa el período de tiempo en el que se recopilaron los datos.
   - Continente:

       Esta variable puede representar el continente al que pertenecen los países involucrados en las transacciones. Por ejemplo, América, Europa, Asia, África, etc.
   - Mes:

       Representa el mes específico en el que se llevaron a cabo las transacciones comerciales. Complementa la información proporcionada por la variable "Años".
   - Año:

       Similar a la variable "Años", indica el año de las transacciones comerciales. Es posible que esta variable contenga la misma información que "Años".
   - Codigopais:

       Esta variable es un código único que identifica a cada país involucrado en las transacciones comerciales. Cada país tiene un código asignado para facilitar su identificación y diferenciación en el análisis.
   - MES:

       Al igual que la variable "Mes", representa el mes en el que se realizaron las transacciones comerciales. Puede contener la misma información que la variable "Mes".

- Peso bruto:

  Indica el peso total de los productos o mercancías involucradas en las transacciones comerciales antes de cualquier ajuste o deducción.

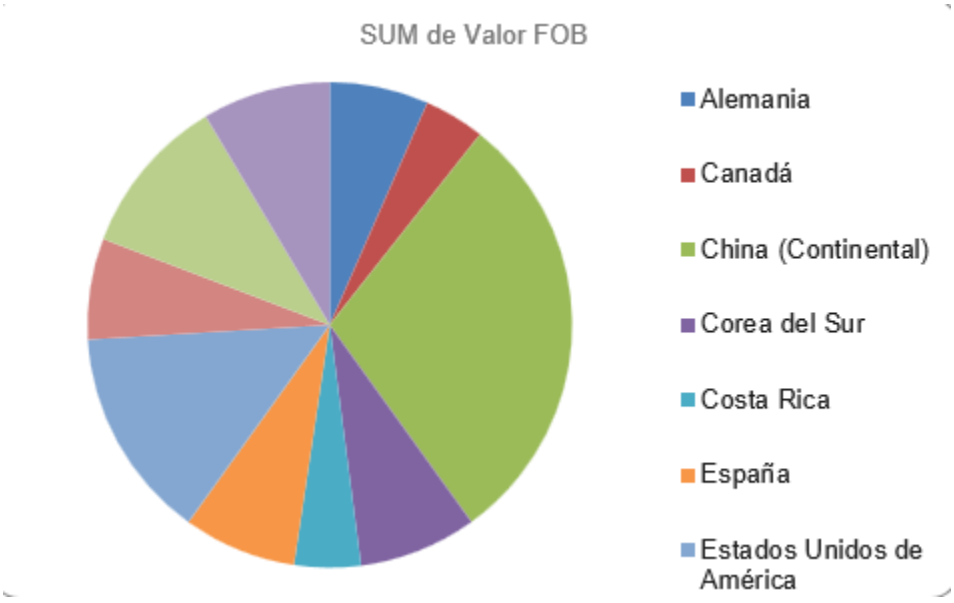
- Peso neto:

  Representa el peso de las mercancías o productos después de deducir cualquier empaque o material de embalaje, es decir, el peso real de la carga.

- Valor FOB:

  Es el valor de las mercancías o productos en términos de su precio de venta libre a bordo (Free On Board), lo que significa que el precio no incluye costos adicionales de transporte o seguro más allá del punto de origen.

Países*	SUM of Valor FOB
Alemania	846706235
Canadá	519239754
China (Continental)	3774642966
Corea del Sur	1010057920
Costa Rica	568371818
España	968516642
Estados Unidos de	1812309244
India	863632511
Japón	1369427183
Países Bajos	1097905168
Total general	**12830809441**



Intervalo de Valor	Fr. Absoluta	Fr. Absoluta acumulativa	Fr. Relativa	Fr. Relativa Acumulada	Suma de los	Media por intervalo	Desviación estándar			
0 - 500	1	1	0,09	0,09	451212023	451212023	0	0	0	**0**
501 - 1000	2	3	0,18	0,27	968516642	484258321	484258321	2,3451E+17	1,1725E+17	**342422343**
1001 - 1500	3	6	0,27	0,55	1010057920	336685973	673371947	4,5343E+17	1,5114E+17	**388771475**
1501 - 2000	2	8	0,18	0,73	1812309244	906154622	906154622	8,2112E+17	4,1056E+17	**640748078**
2000+	3	11	0,27	1,00	3774642966	1258214322	2516428644	6,3324E+18	2,1108E+18	**1452860755**
Total	11	29	1,00	2,64						
Media	2,2	5,8	0,2	0,527						

Distribución de los valores FOB:

Los valores FOB (Free On Board) de las transacciones comerciales se han agrupado en cinco intervalos diferentes.
Concentración en los intervalos de valor bajo y alto:

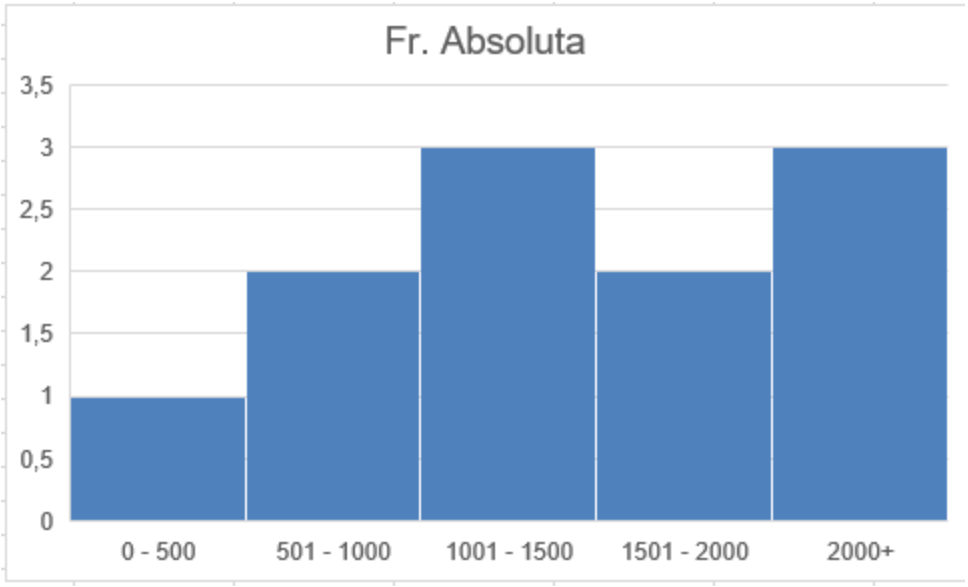
Se observa que el mayor número de transacciones comerciales se encuentra en los intervalos "0 - 500" y "Más de 2000". Esto indica que una parte significativa de las transacciones tiene un valor FOB relativamente bajo o muy alto.
Menos transacciones en el intervalo "1501 - 2000":

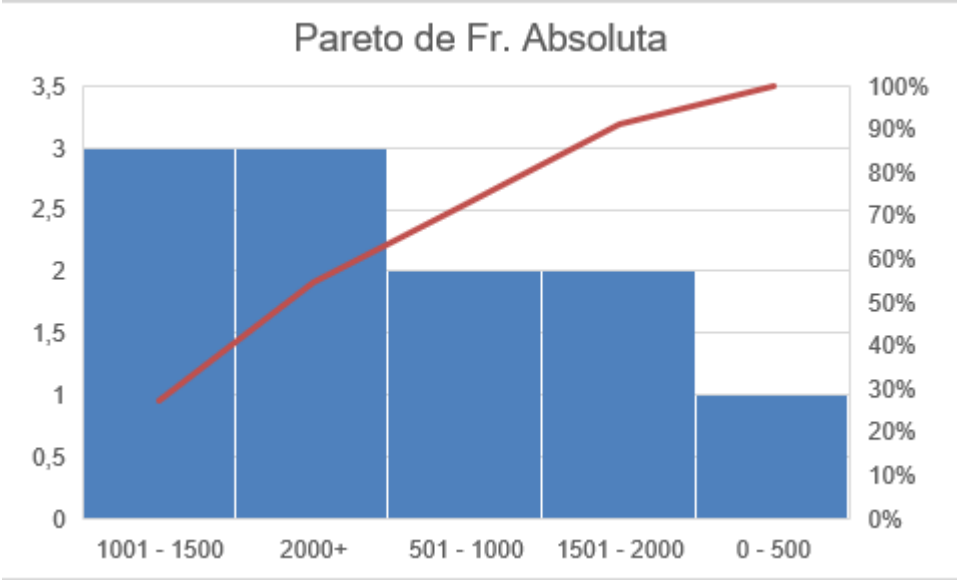
El intervalo "1501 - 2000" tiene el menor número de transacciones, con solo dos registros. Esto indica que hay menos transacciones con valores FOB en este rango.
Distribución acumulativa:

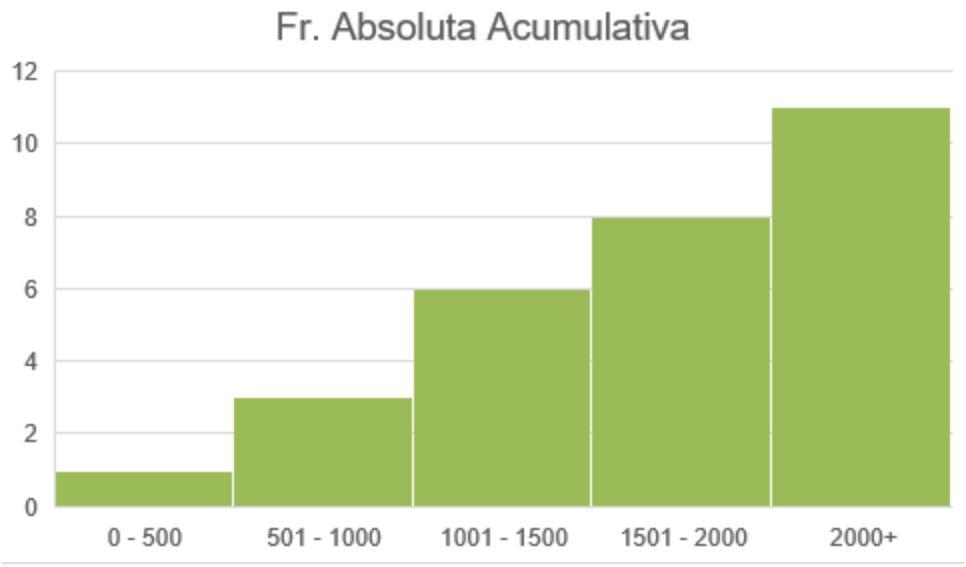
La frecuencia absoluta acumulativa muestra cómo se acumulan los registros a medida que avanzamos en los intervalos. En este caso, aproximadamente el 55% de las transacciones tienen un valor FOB de hasta 1500 millones.
Transacciones con valor FOB alto:

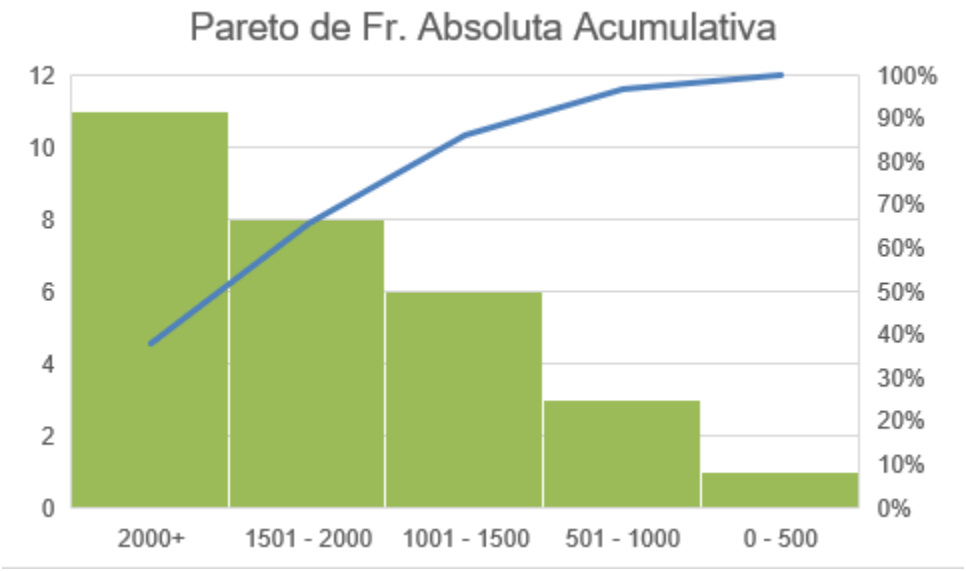
Los intervalos "1001 - 1500" y "Más de 2000" tienen la mayor frecuencia relativa (27% cada uno), lo que indica que hay una cantidad considerable de transacciones con valores FOB más altos.
Ausencia de transacciones en algunos intervalos:

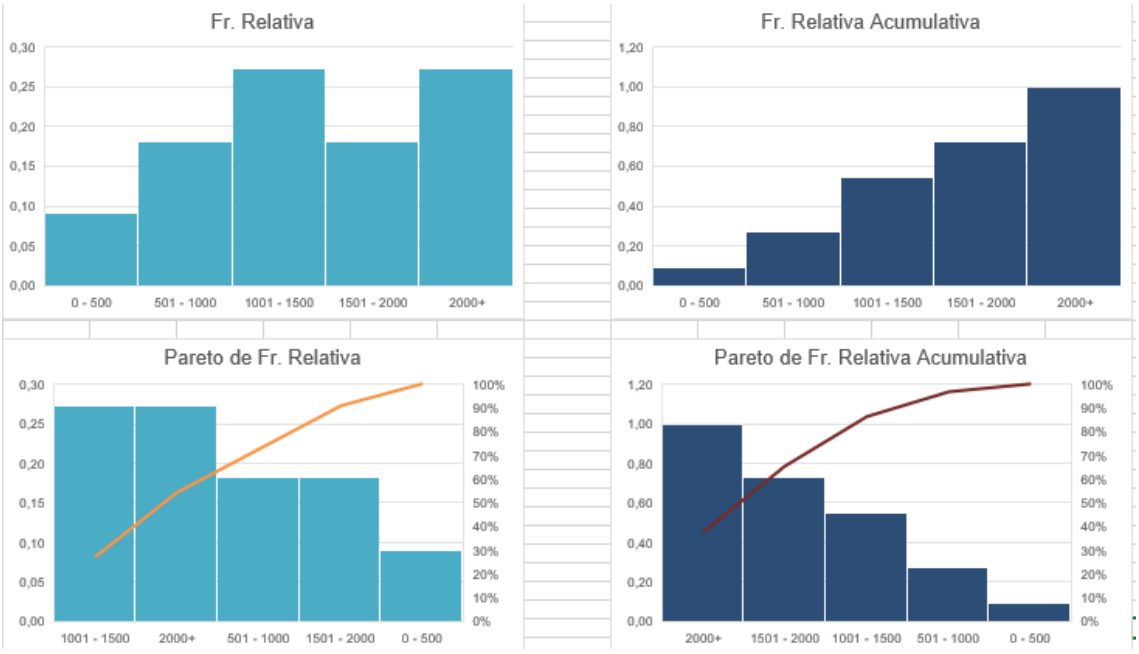
Se puede observar que no hay transacciones registradas en los intervalos "501 - 1000" y "2001 - 2500". Esto podría ser debido a la naturaleza de los datos o a la falta de registros en esos rangos específicos.


Fr. Absoluta
3,5
3
2,5
2
1,5
1
0,5
0
0 - 500
501 - 1000
1001 - 1500
1501 - 2000
2000+


Pareto de Fr. Absoluta
3,5
3
2,5
2
1,5
1
0,5
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1001 - 1500
2000+
501 - 1000
1501 - 2000
0 - 500


Fr. Absoluta Acumulativa
12
10
8
6
4
2
0
0 - 500
501 - 1000
1001 - 1500
1501 - 2000
2000+


Pareto de Fr. Absoluta Acumulativa
12
10
8
6
4
2
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2000+
1501 - 2000
1001 - 1500
501 - 1000
0 - 500



	Intervalo de Valor FOB (en millones)	Fr. Absoluta	\
0	0 - 500	1	
1	501 - 1000	2	
2	1001 - 1500	3	
3	1501 - 2000	2	
4	2000+	3	

	Fr. Absoluta acumulativa	Fr. Relativa	Fr. Relativa Acumulada	\
0	1	0.09	0.09	
1	3	0.18	0.27	
2	6	0.27	0.55	
3	8	0.18	0.73	
4	11	0.27	1.00	

	Valor Representativo	Valor Continuo
0	250.0	(0.0, 500.0]
1	750.5	(501.0, 1000.0]
2	1250.5	(1001.0, 1500.0]
3	1750.5	(1501.0, 2000.0]
4	2000.0	(1501.0, 2000.0]

Transformación de los datos discretos a continuos, usando Python.

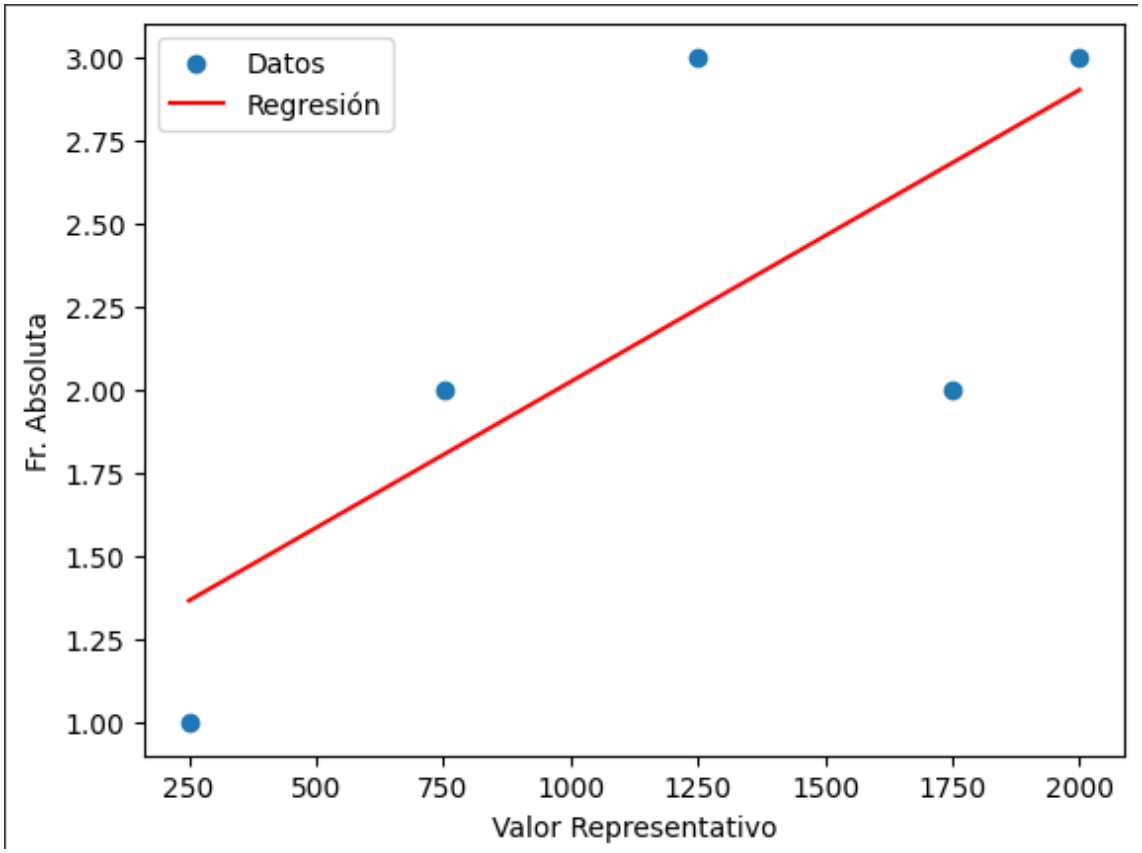
```python
import pandas as pd

# Datos de la tabla de frecuencia con valores continuos
datos = {
    'Intervalo de Valor FOB (en millones)': ['0 - 500', '501 - 1000', '1001 - 1500', '1501 - 2000', '2000+'],
    'Fr. Absoluta': [1, 2, 3, 2, 3],
    'Fr. Absoluta acumulativa': [1, 3, 6, 8, 11],
    'Fr. Relativa': [0.09, 0.18, 0.27, 0.18, 0.27],
    'Fr. Relativa Acumulada': [0.09, 0.27, 0.55, 0.73, 1.00],
    'Valor Representativo': [250, 750.5, 1250.5, 1750.5, 2000],
    'Valor Continuo': [0.09491525423728814, 0.1694915254237288, 0.23728813559322035, 0.2542372881355932, 0.2711864406779661]
}

# Crear el DataFrame con los datos
df = pd.DataFrame(datos)

# Calcular la correlación entre "Valor Continuo" y "Fr. Absoluta"
correlacion = df['Valor Continuo'].corr(df['Fr. Absoluta'])

# Mostrar el resultado
print("Correlación entre Valor Continuo y Fr. Absoluta:", correlacion)
```

```
Correlación entre Valor Continuo y Fr. Absoluta: 0.8537533002961126
```

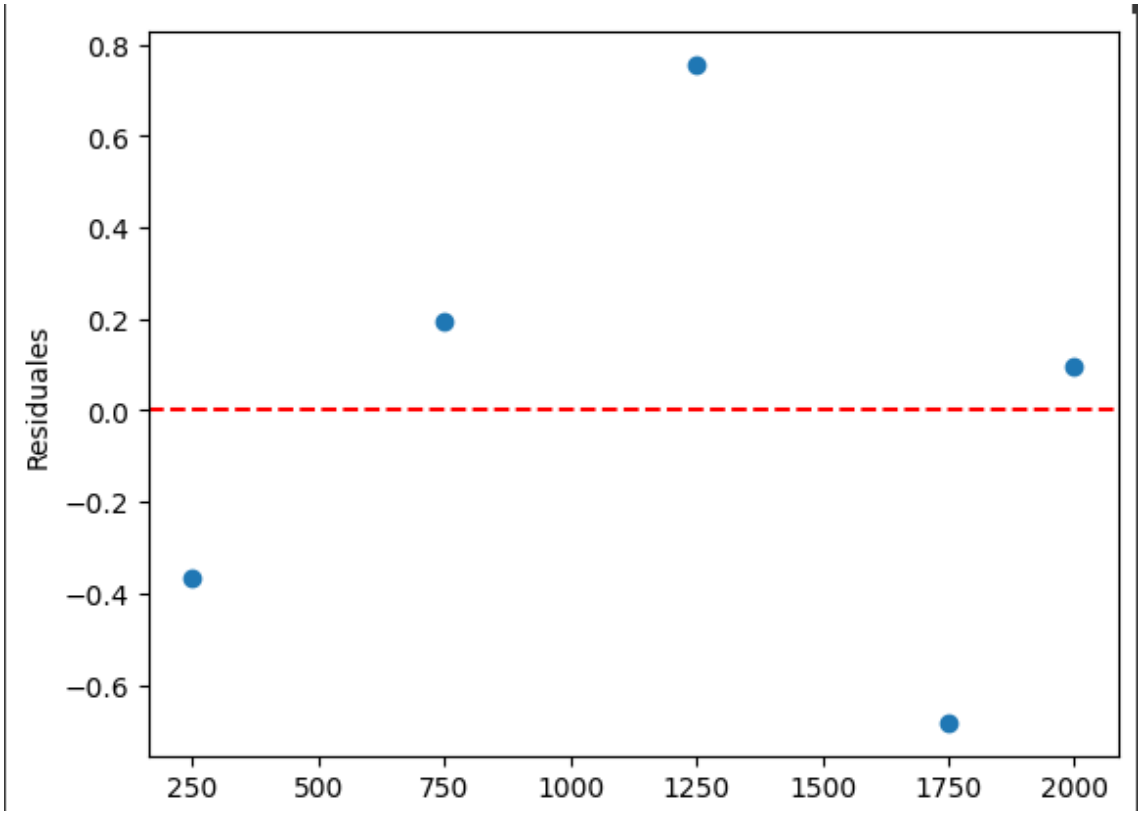
Cálculo de la correlación entre el valor continuo y la fr. Absoluta, realizado en Python.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Fr. Absoluta   R-squared:                       0.565
Model:                            OLS   Adj. R-squared:                  0.419
Method:                 Least Squares   F-statistic:                     3.889
Date:                Wed, 26 Jul 2023   Prob (F-statistic):              0.143
Time:                        03:40:37   Log-Likelihood:                -3.5667
No. Observations:                   5   AIC:                             11.13
Df Residuals:                       3   BIC:                             10.35
Df Model:                           1
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                      1.1460      0.606      1.892      0.155      -0.782       3.074
Valor Representativo       0.0009      0.000      1.972      0.143      -0.001       0.002
==============================================================================
Omnibus:                        nan   Durbin-Watson:                   2.714
Prob(Omnibus):                  nan   Jarque-Bera (JB):                0.266
Skew:                         0.121   Prob(JB):                        0.875
Kurtosis:                     1.896   Cond. No.                     2.89e+03
==============================================================================
```

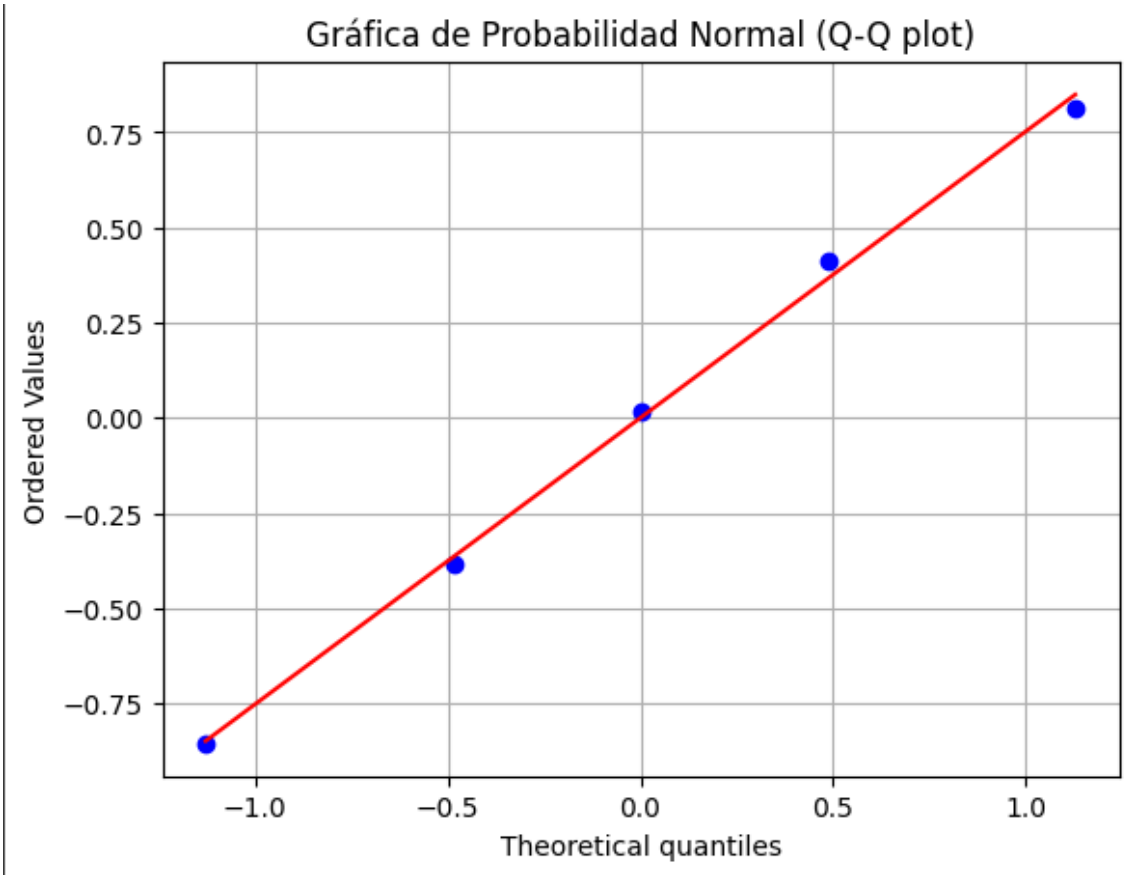
Tabla de regresión lineal y ANOVA



Gráfica de dispersión de puntos con la regresión en línea



Gráfica de residuales



9. Base de datos: FOB en principales exportaciones
   Descripción de la aproximación: Curiosidad
   Descripción de variables:
   Dentro/fuera de País conjunto:

   Esta variable indica si los datos corresponden a transacciones comerciales realizadas dentro del país (dentro de las fronteras del país) o fuera del país (transacciones internacionales o de comercio exterior).

- Grupo:

  Esta variable puede representar una categoría o clasificación específica a la que pertenecen los datos. Por ejemplo, podría ser una clasificación de productos, sectores industriales o cualquier otra agrupación relevante para el análisis.

- Países:

  Esta variable hace referencia a los países involucrados en las transacciones comerciales. Puede haber un país de origen (si se trata de importaciones) y un país de destino (si se trata de exportaciones) dependiendo del contexto de los datos.

- Años:

  Esta variable indica el año en el que se realizaron las transacciones comerciales. Representa el período de tiempo en el que se recopilaron los datos.
- Continente:

  Esta variable puede representar el continente al que pertenecen los países involucrados en las transacciones. Por ejemplo, América, Europa, Asia, África, etc.
- Mes:

  Representa el mes específico en el que se llevaron a cabo las transacciones comerciales. Complementa la información proporcionada por la variable "Años".
- Año:

  Similar a la variable "Años", indica el año de las transacciones comerciales. Es posible que esta variable contenga la misma información que "Años".
- Codigopais:

  Esta variable es un código único que identifica a cada país involucrado en las transacciones comerciales. Cada país tiene un código asignado para facilitar su identificación y diferenciación en el análisis.
- MES:

  Al igual que la variable "Mes", representa el mes en el que se realizaron las transacciones comerciales. Puede contener la misma información que la variable "Mes".
- Peso bruto:

  Indica el peso total de los productos o mercancías involucradas en las transacciones comerciales antes de cualquier ajuste o deducción.
- Peso neto:

  Representa el peso de las mercancías o productos después de deducir cualquier empaque o material de embalaje, es decir, el peso real de la carga.
- Valor FOB:

  Es el valor de las mercancías o productos en términos de su precio de venta libre a bordo (Free On Board), lo que significa que el precio no incluye costos adicionales de transporte o seguro más allá del punto de origen.

Intervalos		Frecuencia	Frecuencia Abs.	Frecuencia Abs. Acumulativa	Frecuencia Relativa	Frecuencia Rel. Acumulativa	Media por intervalo
2010-2011	1263577127	7952	7952	7952	9.71	9.71	158900.544
2012-2013	295962263	2106	2106	10058	2.57	12.29	140532.888
2014-2015	1950988733	14157	14157	24215	17.29	29.58	137810.887
2016-2017	4282764835	6157	6157	30372	7.52	37.10	695592.794
2018-2019	3143911407	21828	21828	52200	26.67	63.77	144031.125
2020-2021	6506960921	22385	22385	74585	27.35	91.12	290683.981
2022-2023	685797637	7272	7272	81857	8.88	100	94306.6057
Total	1.813E+10	81857	81857	281239	100	343.573549	1661858.83
Media	2589994703	11693.8571	11693.8571	40177	14.286	49.082	237408.404

Tabla de frecuencias y media

Distribución de datos por intervalo de años:

Los datos se han agrupado en diferentes intervalos de años, desde 2010 hasta 2023.
Distribución de la frecuencia en los intervalos:

Se observa que los intervalos "2010-2011" y "2020-2021" tienen la mayor frecuencia, con 7952 y 22385 observaciones respectivamente. Esto sugiere que hay una concentración significativa de datos en esos períodos.
Crecimiento en la frecuencia de datos:

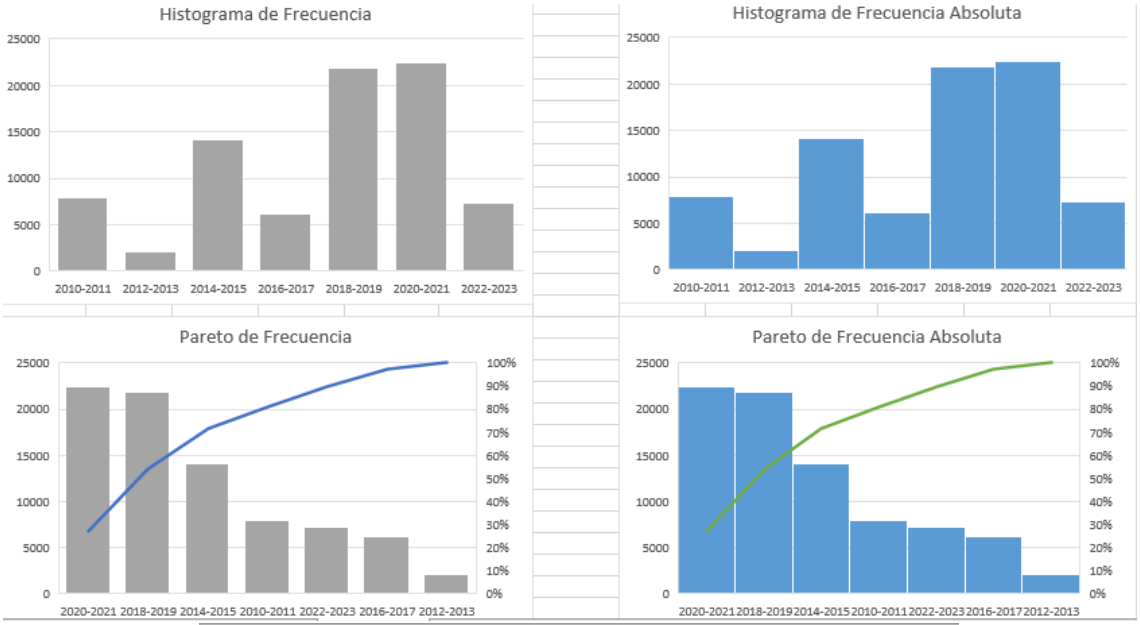
La frecuencia absoluta acumulativa muestra cómo aumenta el número de observaciones a medida que avanzamos en los intervalos. Se puede observar que la frecuencia de datos va aumentando gradualmente a lo largo de los años.
Media por intervalo de años:

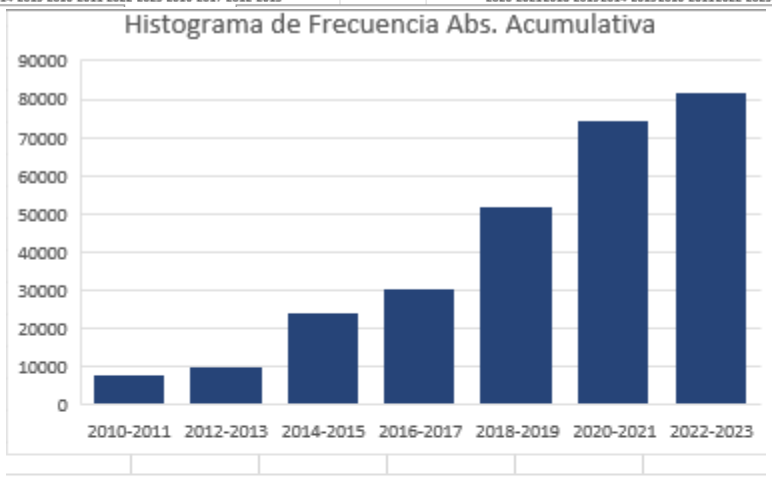
La columna "Media por intervalo" indica el promedio de datos por intervalo. Los años 2018-2019 tienen el promedio más alto con aproximadamente 144031 observaciones, mientras que los años 2012-2013 tienen el promedio más bajo con alrededor de 140532 observaciones.
Distribución acumulativa:

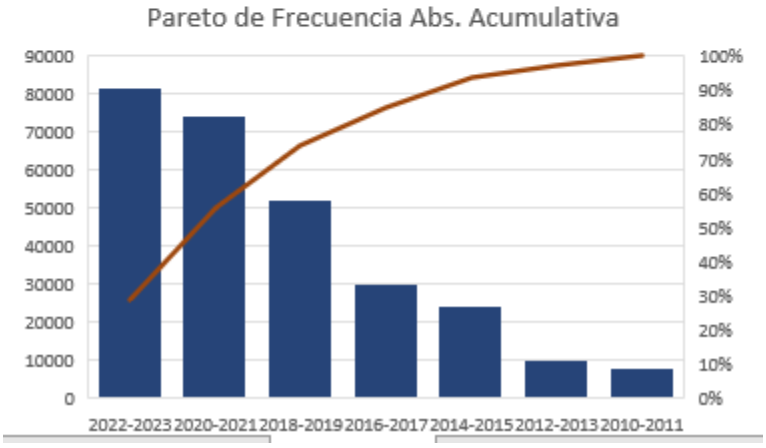
La frecuencia relativa acumulativa muestra cómo se acumulan las observaciones a medida que avanzamos en los intervalos. Cerca del 91.12% de los datos se encuentran en los intervalos hasta 2021 (incluido).
Tendencia de crecimiento a lo largo del tiempo:

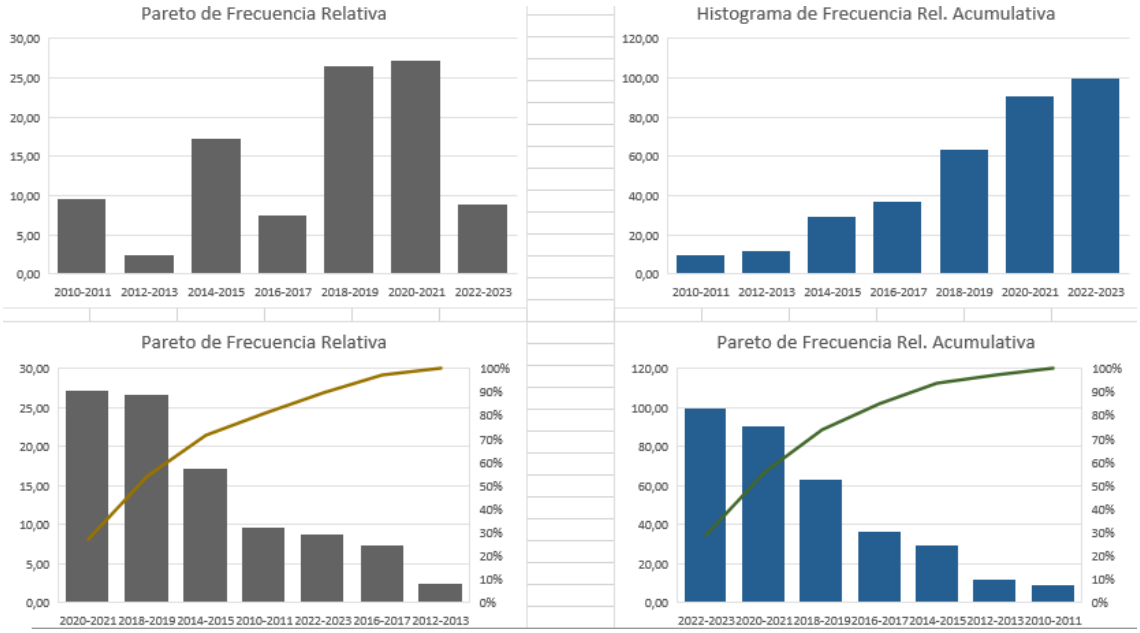
En general, se puede observar que el número de observaciones ha ido aumentando a lo largo del tiempo, con intervalos más recientes teniendo una mayor cantidad de datos.

Gráfico de los datos en bruto.


Histograma de Frecuencia
25000
20000
15000
10000
5000
0
2010-2011 2012-2013 2014-2015 2016-2017 2018-2019 2020-2021 2022-2023
Histograma de Frecuencia Absoluta
25000
20000
15000
10000
5000
0
2010-2011 2012-2013 2014-2015 2016-2017 2018-2019 2020-2021 2022-2023
Pareto de Frecuencia
25000
20000
15000
10000
5000
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2020-2021 2018-2019 2014-2015 2010-2011 2022-2023 2016-2017 2012-2013
Pareto de Frecuencia Absoluta
25000
20000
15000
10000
5000
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2020-2021 2018-2019 2014-2015 2010-2011 2022-2023 2016-2017 2012-2013

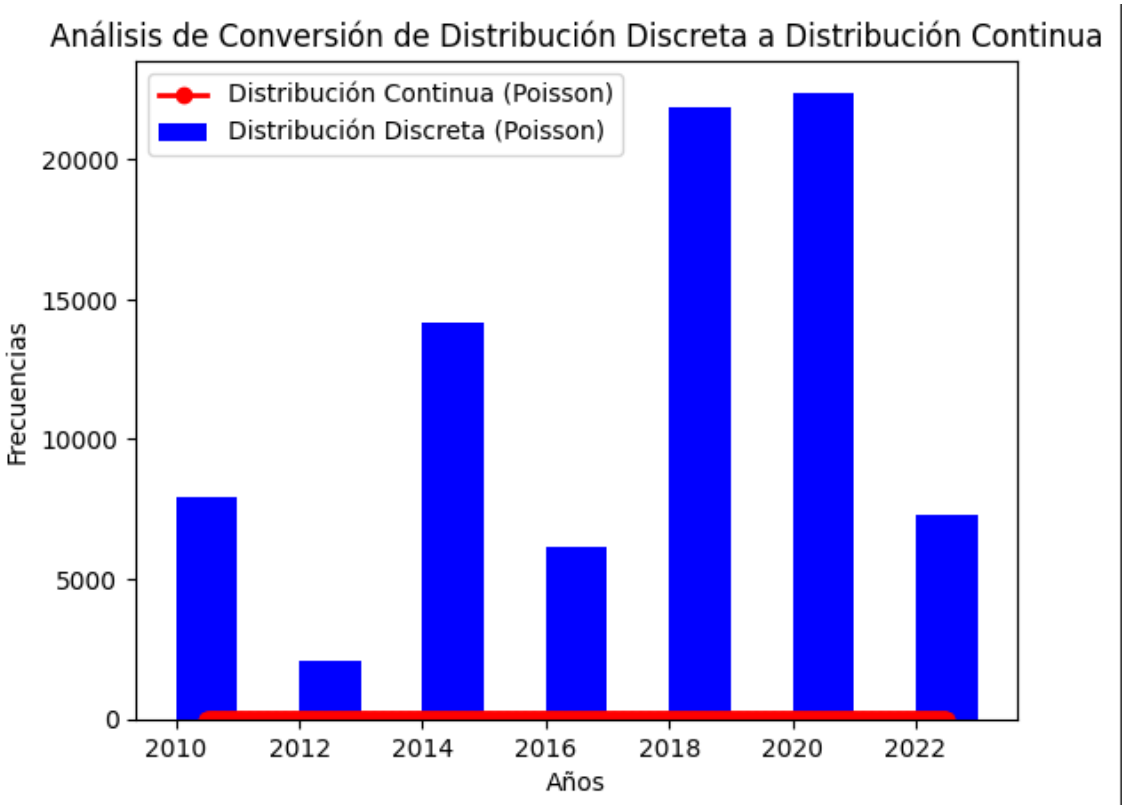

Histograma de Frecuencia Abs. Acumulativa
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
2010-2011 2012-2013 2014-2015 2016-2017 2018-2019 2020-2021 2022-2023


Pareto de Frecuencia Abs. Acumulativa
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2022-2023 2020-2021 2018-2019 2016-2017 2014-2015 2012-2013 2010-2011



Intervalos	Frecuencia	Frecuencia Abs.	Frecuencia Abs. Acumulativa	Frecuencia Relativa	Frecuencia Rel. Acumulativa	Media por intervalo		Desviación estándar		
2010-2011	1263577127	7952	7952	7952	9,71	9,71	158900,544	1263418226,456	1596225614940870000	2,00733E+14	14168013,2
2012-2013	295962263	2106	2106	10058	2,57	12,29	140532,888	295821730,112	87510496006212000	4,15529E+13	6446157,16
2014-2015	1950988733	14157	14157	24215	17,29	29,58	137810,887	1950850922,113	3805819320307560000	2,6883E+14	16396021,1
2016-2017	4282764835	6157	6157	30372	7,52	37,10	695592,794	4282069242,206	18336116995043400000	2,97809E+15	54571904,3
2018-2019	3143911407	21828	21828	52200	26,67	63,77	144031,125	3143767375,875	9883273313612960000	4,5278E+14	21278618,6
2020-2021	6506960921	22385	22385	74585	27,35	91,12	290683,981	6506670237,019	42336757573305200000	1,8913E+15	43489082,7
2022-2023	685797637	7272	7272	81857	8,88	100	94306,6057	685703330,394	470189057313769000	6,46575E+13	8040986,33
Total	1,813E+10	81857	81857	281239	100	343,573549	1661858,83	18128301064	7,65159E+19	5,89795E+15	164390783
Media	2589994703	11693,8571	11693,8571	40177	14,286	49,082	237408,404	2589757295	1,09308E+19	8,42564E+14	23484397,6

Tabla de la desviación estándar



Gráfico

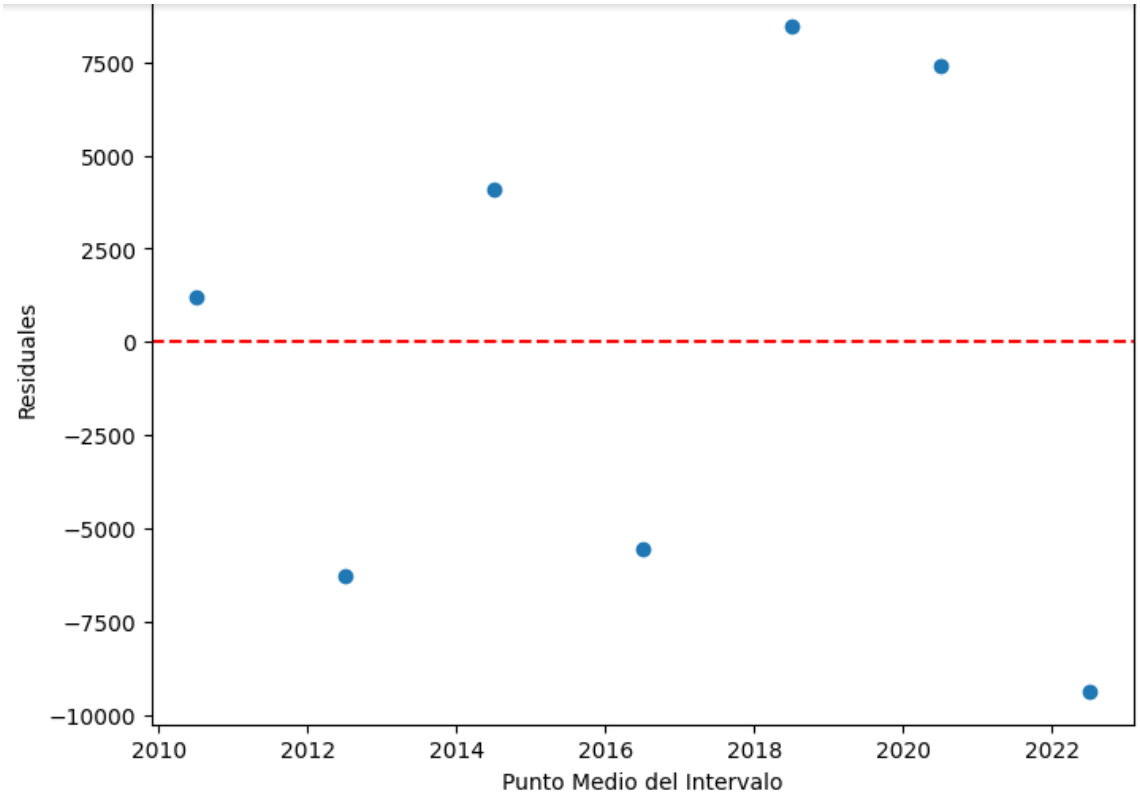
```python
import numpy as np

# Datos de las variables
intervalos = [2010, 2012, 2014, 2016, 2018, 2020, 2022]
frecuencia = [7952, 2106, 14157, 6157, 21828, 22385, 7272]

# Calcula el coeficiente de correlación de Pearson
correlation_coef = np.corrcoef(intervalos, frecuencia)[0, 1]

print("Coeficiente de correlación de Pearson:", correlation_coef)
```

```
Coeficiente de correlación de Pearson: 0.44825136356227563
```

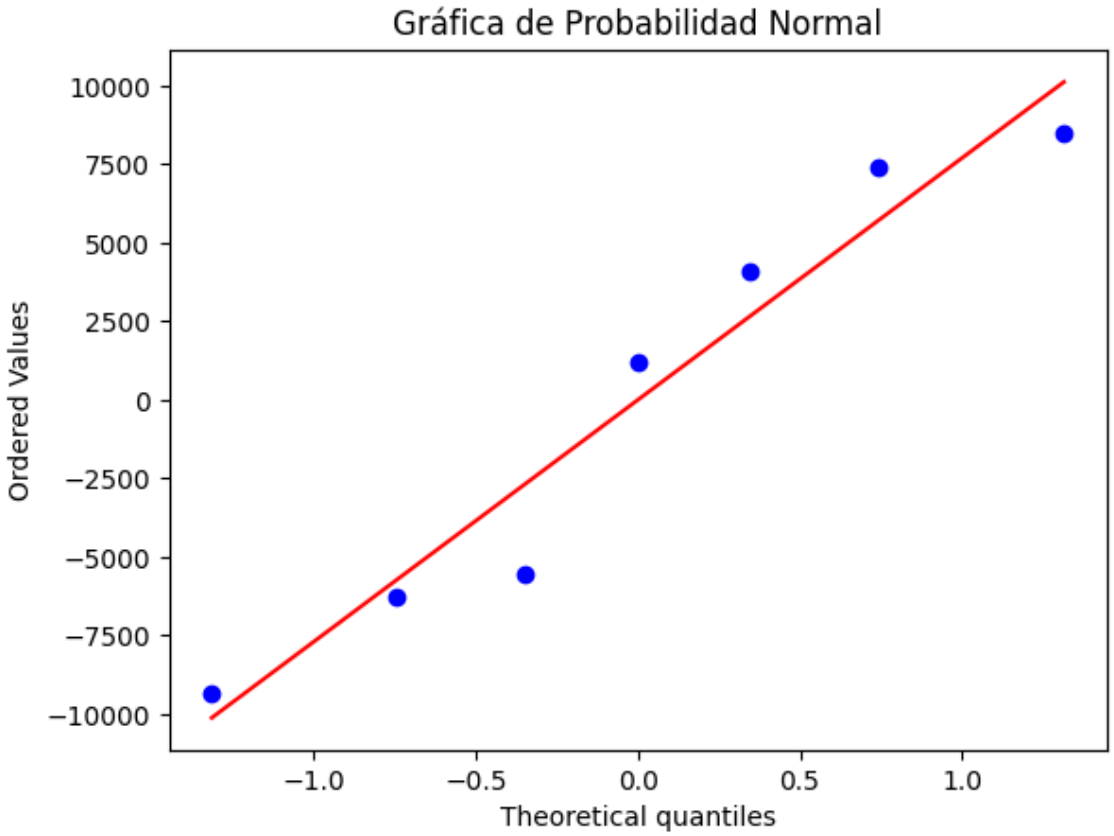
Resultado de la correlación usando Python.

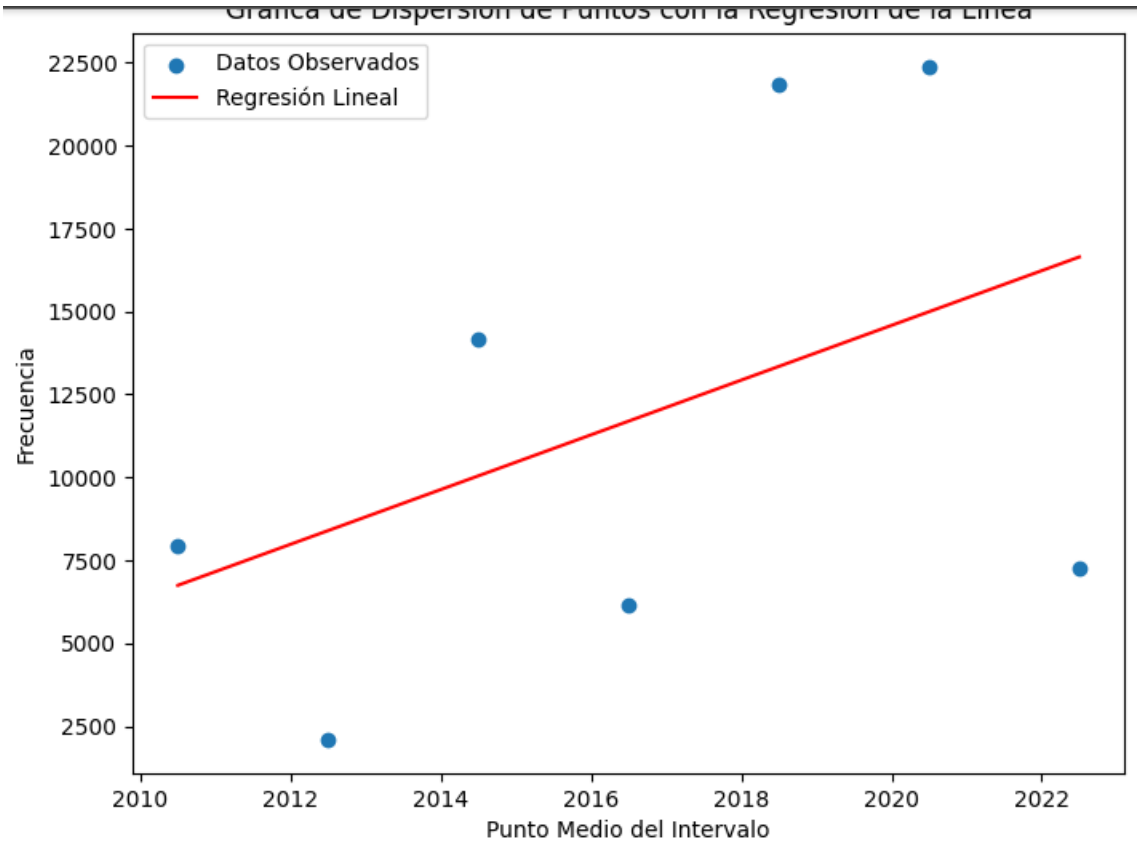
```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Frecuencia   R-squared:                       0.201
Model:                            OLS   Adj. R-squared:                  0.041
Method:                 Least Squares   F-statistic:                     1.257
Date:                Tue, 25 Jul 2023   Prob (F-statistic):              0.313
Time:                        23:45:32   Log-Likelihood:                -71.474
No. Observations:                   7   AIC:                             146.9
Df Residuals:                       5   BIC:                             146.8
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.652e+06   1.48e+06     -1.113      0.316   -5.46e+06    2.16e+06
Punto_Medio   824.8036    735.591      1.121      0.313   -1066.094    2715.701
==============================================================================
Omnibus:                        nan   Durbin-Watson:                   2.430
Prob(Omnibus):                  nan   Jarque-Bera (JB):                0.732
Skew:                        -0.078   Prob(JB):                        0.693
Kurtosis:                     1.423   Cond. No.                     1.02e+06
==============================================================================
```

Tabla de regresión lineal y ANOVA



Gráfica de residuales


Gráfica de Probabilidad Normal
Ordered Values
Theoretical quantiles
10000
7500
5000
2500
0
−2500
−5000
−7500
−10000
−1.0
−0.5
0.0
0.5
1.0



Gráfica de dispersión de puntos con la regresión de la línea

10. Base de datos: Aporte absoluto de extranjeros en nuestro país

Aproximación por curiosidad

Años:

Esta variable indica el año al que corresponden los datos. Representa el período de tiempo en el que se realizaron las observaciones o mediciones.

Relación - Año:

Esta variable puede ser una relación o proporción con respecto al año específico al que está asociada. Puede ser una relación respecto al año anterior o a otro año de referencia.

Año:

Similar a la variable "Años", indica el año al que se refiere la variable "Relación - Año". Es posible que esta variable contenga la misma información que "Años".

Aporte Absoluto:

Es el aporte o contribución en términos absolutos, sin considerar direcciones o signos. Representa una cantidad específica en unidades absolutas.

Aporte Absoluto Constante:

Esta variable puede representar el aporte absoluto ajustado o calculado utilizando un valor constante como referencia. El valor constante puede ser un parámetro o una cifra de referencia establecida.

Aporte Absoluto Corriente:

Similar a la variable "Aporte Absoluto Constante", pero en este caso el cálculo se realiza utilizando un valor corriente o actual como referencia.

Valor Constante:

Representa el valor calculado o medido utilizando un valor constante como factor de ajuste o comparación. Puede ser una cifra que se mantiene constante en todas las observaciones.

Valor Corriente:

Similar a la variable "Valor Constante", pero en este caso el cálculo se realiza utilizando un valor corriente o actual como factor de ajuste o comparación.

Valores:

Esta variable puede contener los valores o datos específicos que se están analizando o comparando. Representa las observaciones o mediciones de interés.

Gráfico del análisis de los datos en bruto

Intervalos	Suma	Frecuencia	Fre. Absoluti	Fre. Abs. Acu	Fre. Relativa	Fre. Rel. Acu	Media por intervalo
2007-2009	8142	3	3	3	23,077	23,077	2714
2010-2012	16158	3	3	6	23,077	46,154	5386
2013-2015	12307	3	3	9	23,077	69,231	4102
2016-2018	9076	3	3	12	23,077	92,308	3025
2019-2020	-13007	1	1	13	7,692	100	-13007
Totales	32676	13	13	43	100	330,769	2220,667
Media	6535,20	2,60	2,60	8,60	20,00	66,15	444,13

Tabla de frecuencias y media

Distribución de datos por intervalo de años:

Los datos se han agrupado en cinco intervalos de años, cada uno abarcando un período de tres años.
Resumen de los datos por intervalo:

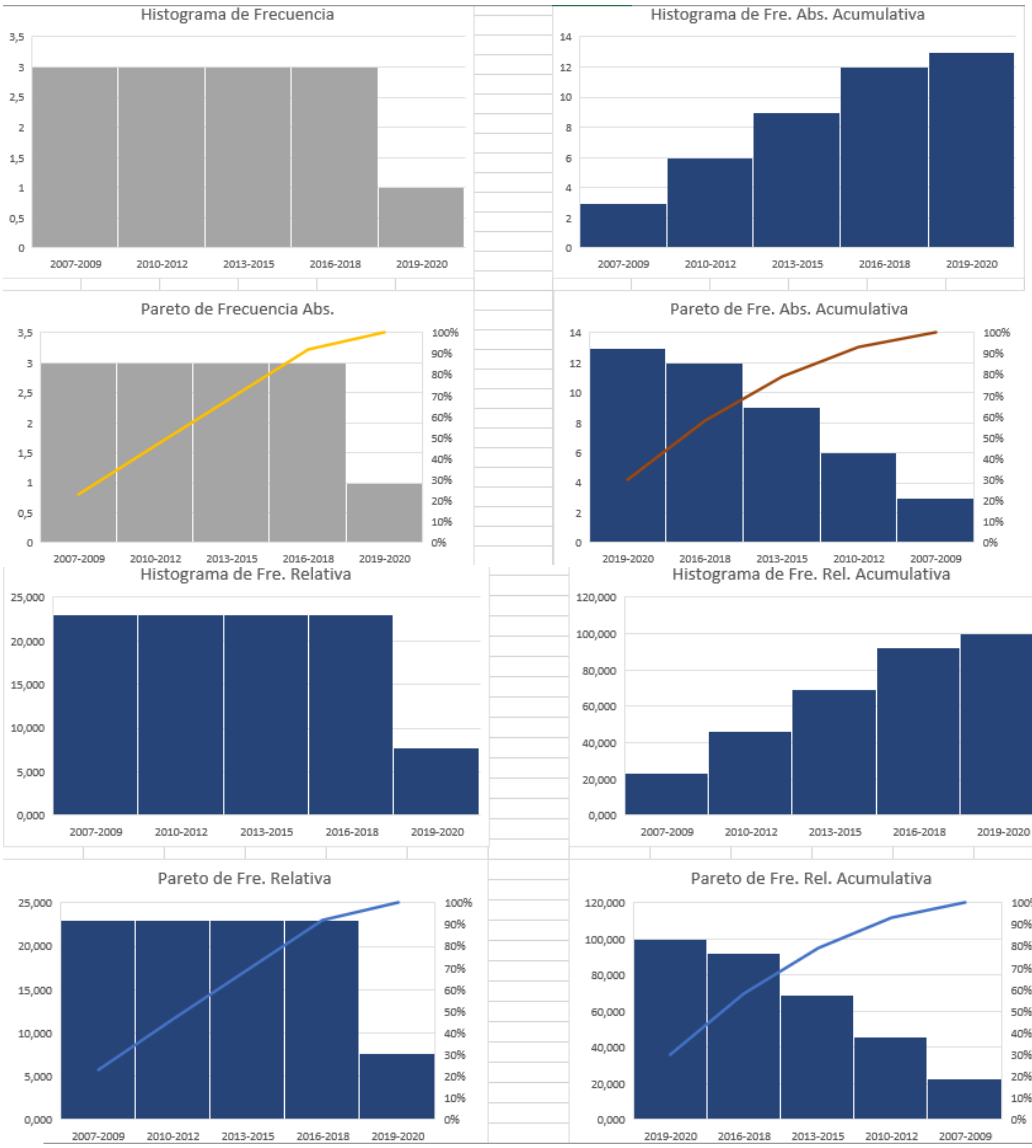
La tabla proporciona información resumida sobre la suma total de los valores (Suma) y el número de datos (Frecuencia) dentro de cada intervalo.
Distribución acumulativa de frecuencia:

La frecuencia absoluta acumulada muestra cómo se acumulan las observaciones a medida que avanzamos en los intervalos. La frecuencia relativa acumulada muestra la proporción acumulada de los datos en cada intervalo.
Media por intervalo:

La columna "Media por intervalo" indica el promedio de los datos dentro de cada intervalo. Los intervalos "2010-2012" y "2013-2015" tienen medias más altas en comparación con los otros intervalos.
Media general de los datos:

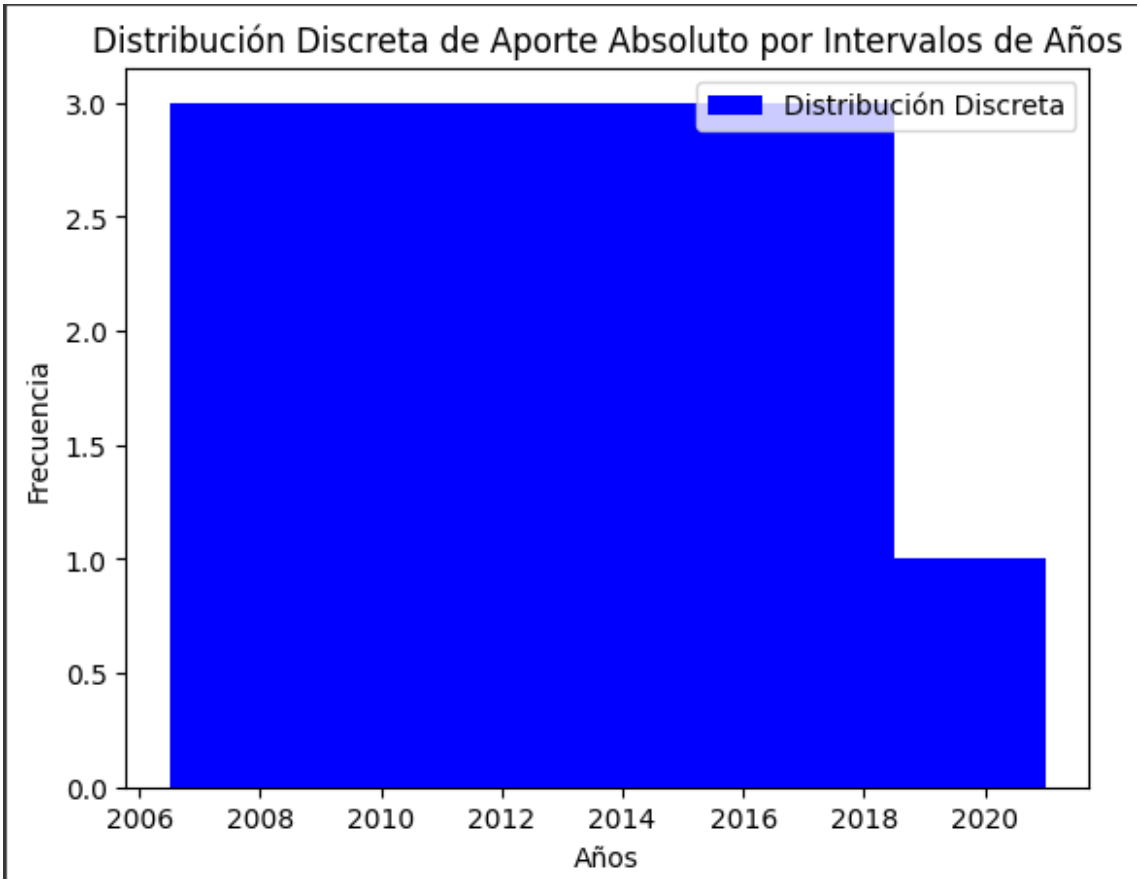
La "Media" general de los datos es de 6535.20, lo que representa el promedio general de todos los valores en la tabla. Esto proporciona una idea del valor promedio que se puede esperar en cada intervalo.
Anomalía en el intervalo "2019-2020":

Se observa una anomalía en el intervalo "2019-2020" con una suma de -13007, lo que indica que la suma de los datos es negativa en ese período.



Intervalos	Suma	Frecuencia	Fre. Absolut	Fre. Abs. Acu	Fre. Relativa	Fre. Rel. Acu	Media por in		Desviación estándar		
2007-2009	8142	3	3	3	23,077	23,077	2714	5428	29463184	9821061,33	3133,85726
2010-2012	16158	3	3	6	23,077	46,154	5386	10772	116035984	38678661,3	6219,2171
2013-2015	12307	3	3	9	23,077	69,231	4102	8204,67	67316555,1	22438851,7	4736,96651
2016-2018	9076	3	3	12	23,077	92,308	3025	6050,67	36610567,1	12203522,4	3493,35403
2019-2020	-13007	1	1	13	7,692	100	-13007	0	0	0	0
Totales	32676	13	13	43	100	330,769	2220,667				
Media	6535,20	2,60	2,60	8,60	20,00	66,15	444,13				

Tabla de frecuencias, media y desviación estándar


Distribución Discreta de Aporte Absoluto por Intervalos de Años
Distribución Discreta
Frecuencia
Años
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2006
2008
2010
2012
2014
2016
2018
2020

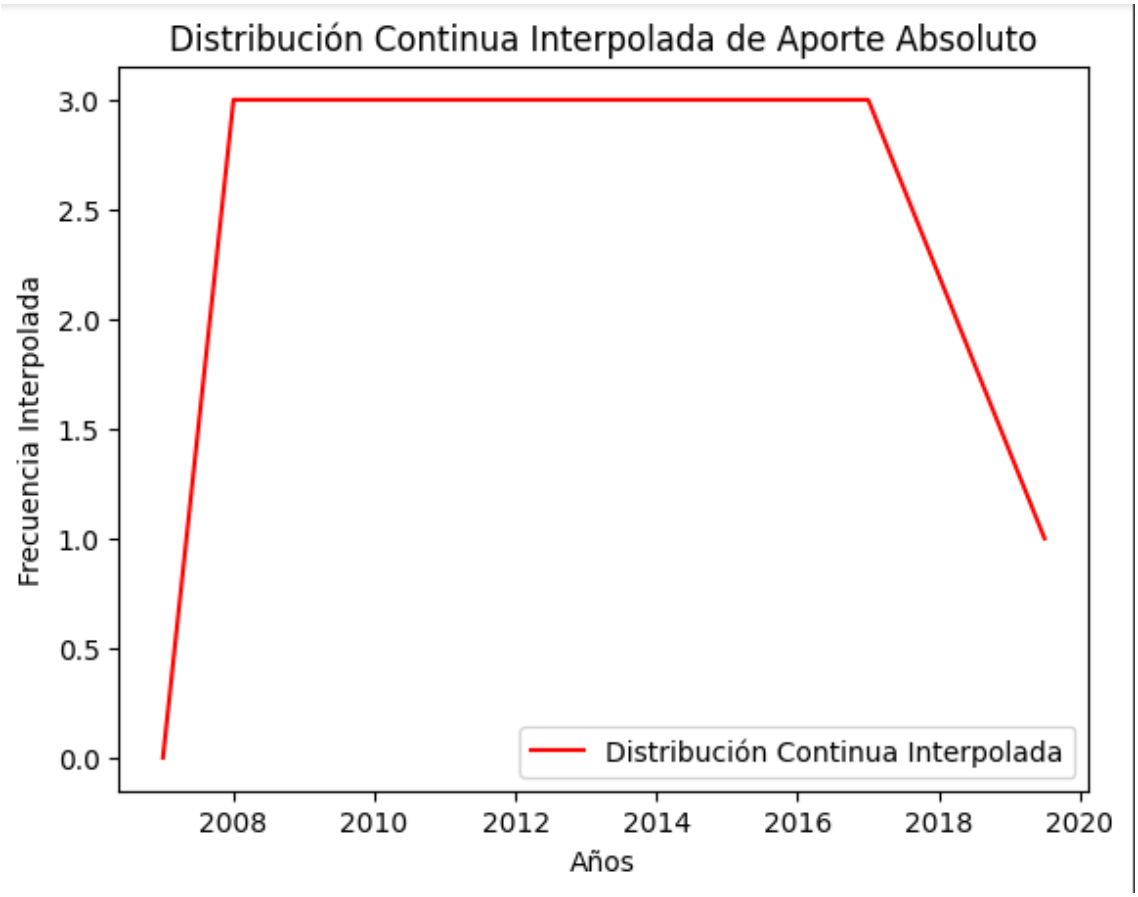
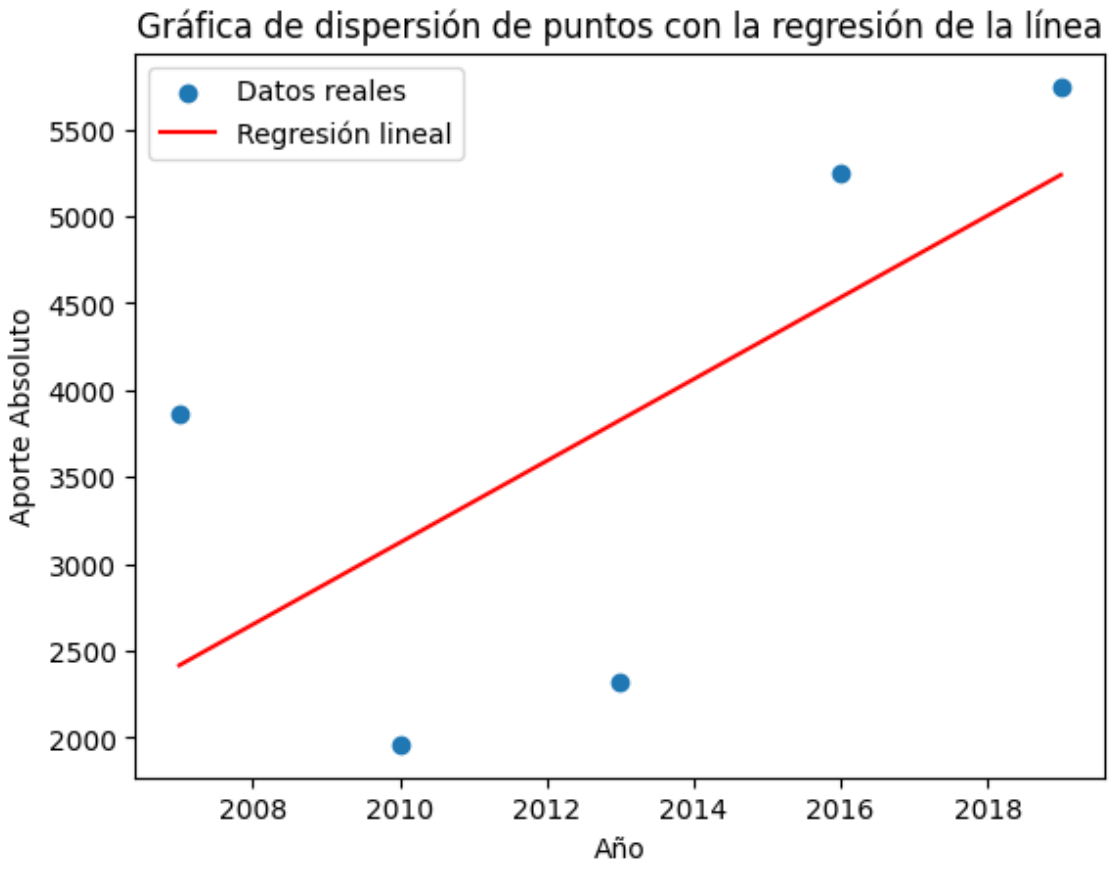
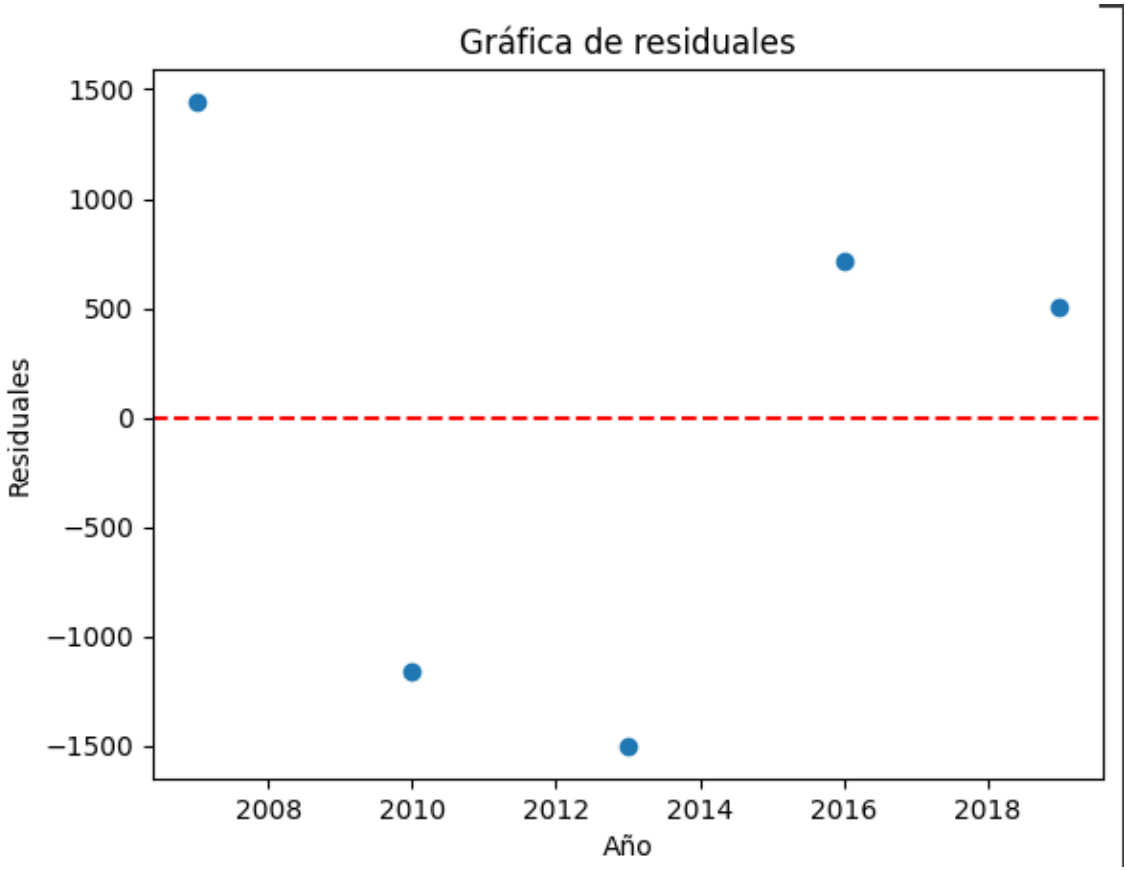


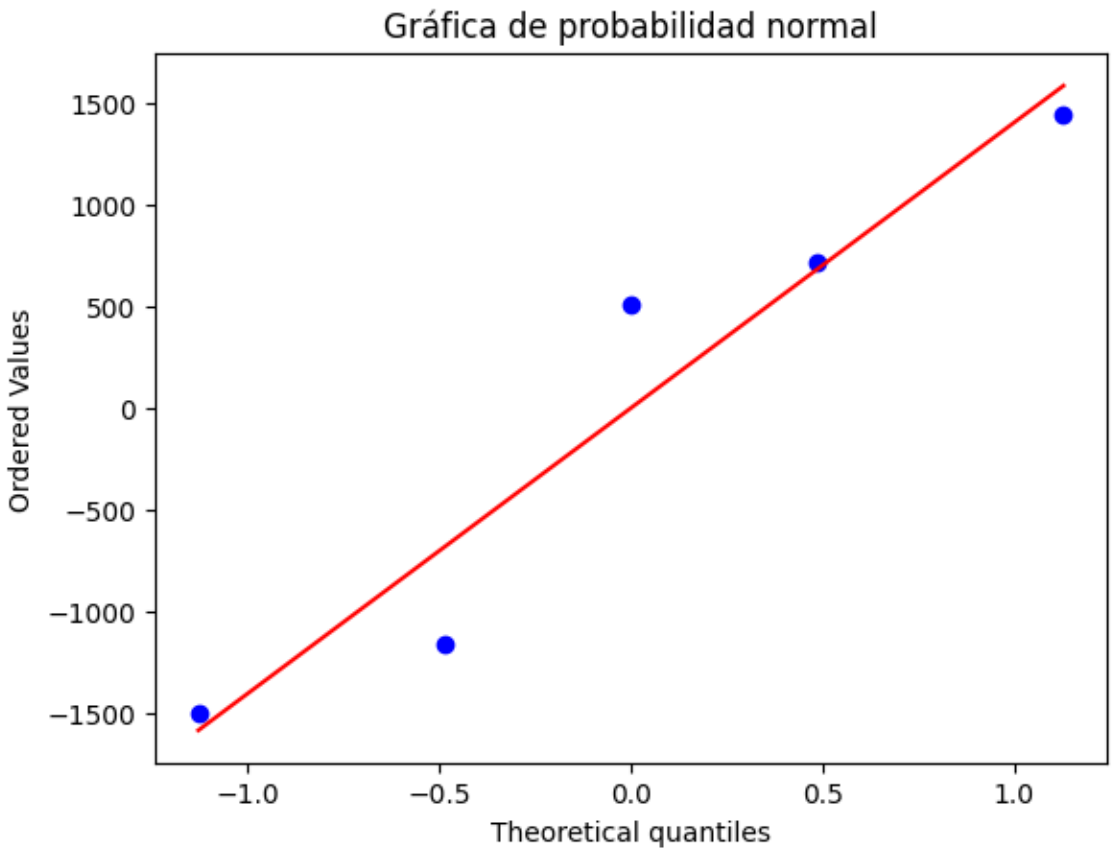
Gráfico de la distribución discreta a continúa interpolada en Python.

```
[ ]  import numpy as np

     # Datos de Año y Aporte Absoluto
     años = [2007, 2010, 2013, 2016, 2019]
     aporte_absoluto = [3859.9, 1960.7, 2323.7, 5245.9, 5743.5]

     # Calcula la matriz de correlación
     correlacion_matrix = np.corrcoef(años, aporte_absoluto)

     # El coeficiente de correlación está en la posición (0, 1) o (1, 0) de la matriz
     coef_correlacion = correlacion_matrix[0, 1]

     print("Coeficiente de correlación de Pearson:", coef_correlacion)

     Coeficiente de correlación de Pearson: 0.659641207721966
```

Resultado de la correlación en Python.


Gráfica de dispersión de puntos con la regresión de la línea
Datos reales
Regresión lineal
Aporte Absoluto
Año
5500
5000
4500
4000
3500
3000
2500
2000
2008
2010
2012
2014
2016
2018


Gráfica de residuales
1500
1000
500
0
−500
−1000
−1500
Residuales
2008
2010
2012
2014
2016
2018
Año



OLS Regression Results

Dep. Variable:	Aporte Absoluto	R-squared:	0.435
Model:	OLS	Adj. R-squared:	0.247
Method:	Least Squares	F-statistic:	2.311
Date:	Wed, 26 Jul 2023	Prob (F-statistic):	0.226
Time:	02:30:40	Log-Likelihood:	-42.273
No. Observations:	5	AIC:	88.55
Df Residuals:	3	BIC:	87.76
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.694e+05	3.11e+05	-1.508	0.229	-1.46e+06	5.21e+05
Año	235.0800	154.640	1.520	0.226	-257.055	727.215

Omnibus:	nan	Durbin-Watson:	1.837
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.573
Skew:	-0.199	Prob(JB):	0.751
Kurtosis:	1.390	Cond. No.	9.55e+05