Universidad Católica Andrés Bello
Facultad de Ingeniería
Escuela de Ingeniería Informática
Análisis de datos

Proyecto Práctico 1

Alumno:
Tadeo Vázquez, CI:27.752.958
Santiago Figueroa, CI: 28.328.058

Caracas, 10 de Mayo de 2023

Realice una investigación sobre Análisis de Datos aplicado al Sector Educación destacando lo siguiente: técnicas más frecuentes e ilustre con dos ejemplos.

Análisis de Datos aplicado al Sector Educación

En el sector educativo, el análisis de datos se utiliza para comprender mejor los procesos de enseñanza y aprendizaje, identificar áreas de mejora y tomar decisiones informadas para mejorar los resultados educativos. A continuación, se presentan algunas de las técnicas más utilizadas en análisis de datos en educación.

Análisis descriptivo: esta técnica implica la descripción de datos a través de medidas estadísticas como la media, la mediana y la desviación estándar. El análisis descriptivo se utiliza a menudo para resumir datos de pruebas estandarizadas y otros resultados de evaluación. Por ejemplo, en un estudio realizado por McLeod, Weininger y Martorell (2020), se utilizó el análisis descriptivo para examinar los resultados de una encuesta que evaluaba las actitudes de los estudiantes hacia la educación en línea durante la pandemia de COVID-19.

Análisis de regresión: el análisis de regresión se utiliza para identificar la relación entre una variable dependiente y una o más variables independientes. Esta técnica se utiliza a menudo para examinar cómo ciertos factores pueden afectar el rendimiento de los estudiantes. Por ejemplo, en un estudio realizado por Bryner y Ye (2020), se utilizó el análisis de regresión para examinar cómo la calidad de la enseñanza afectaba el rendimiento académico de los estudiantes universitarios.

Análisis de conglomerados: esta técnica se utiliza para identificar grupos de estudiantes que tienen características similares. El análisis de conglomerados puede ser útil para identificar grupos de estudiantes que pueden necesitar intervenciones específicas. Por ejemplo, en un estudio realizado por Quek, Wong y Lin (2020), se utilizó el análisis de conglomerados para identificar grupos de estudiantes que tenían diferentes patrones de comportamiento relacionados con la tecnología.

Análisis de redes sociales: se utiliza para examinar las interacciones entre los estudiantes y otros miembros de la comunidad educativa, como los profesores y los administradores escolares. El análisis de redes sociales puede ser útil para comprender mejor las dinámicas sociales en la escuela y cómo estas dinámicas pueden afectar el rendimiento académico de los estudiantes. Por ejemplo, en un estudio realizado por Spillane y Hallett (2019), se utilizó el análisis de redes sociales para examinar cómo la colaboración entre los profesores puede afectar el rendimiento académico de los estudiantes.

Un ejemplo de la aplicación de la analítica de aprendizaje es el proyecto desarrollado por la Universidad de Harvard, donde se utilizó esta herramienta para identificar patrones de comportamiento de los estudiantes y mejorar la eficacia de los cursos en línea. Se analizaron los datos generados por los estudiantes, como el tiempo que pasaban en el curso, las actividades que realizaban y las calificaciones obtenidas, para identificar patrones y tendencias que permitieran mejorar el diseño del curso y la experiencia de aprendizaje de los estudiantes

Por otro lado, un ejemplo de la aplicación del análisis de datos en el aula es el proyecto desarrollado por la Universidad de Stanford, donde se utilizó el análisis de datos para identificar patrones de comportamiento

de los estudiantes y mejorar la eficacia de los cursos presenciales. Se analizaron los datos generados por los profesores y los estudiantes durante el proceso de enseñanza-aprendizaje, como las calificaciones, las interacciones en el aula y las respuestas a las preguntas, para identificar patrones y tendencias que permitieran mejorar la eficacia del curso

En conclusión, la aplicación de técnicas de análisis de datos en el sector educativo puede mejorar la calidad de la educación. La analítica de aprendizaje y el análisis de datos en el aula son técnicas comunes que pueden ser aplicadas para mejorar el proceso de enseñanza-aprendizaje, esta sirve como una herramienta valiosa para el sector educativo, ya que puede ayudar a identificar áreas de mejora y tomar decisiones informadas para mejorar los resultados educativos. Las técnicas más utilizadas incluyen el análisis descriptivo, el análisis de regresión, el análisis de conglomerados y el análisis de redes sociales.

Usando los conjuntos de datos `calificaciones` y `paises`, realice para cada uno un análisis de componentes principales y establezca conclusiones.

Para Calificaciones:

Primero se reduce la dimensionalidad de un conjunto de datos compuesto por las calificaciones de 20 estudiantes en 6 asignaturas diferentes. El objetivo de este análisis fue encontrar patrones y relaciones ocultas entre las variables originales.

Para la realización de los gráficos se utilizó Python como herramienta de apoyo usando las librerías : pandas, numpy, sklearn y matplotlib.

Primero se realizó un gráfico de caja bigote para ver la distribución de los datos en términos de las medidas relativas de posición

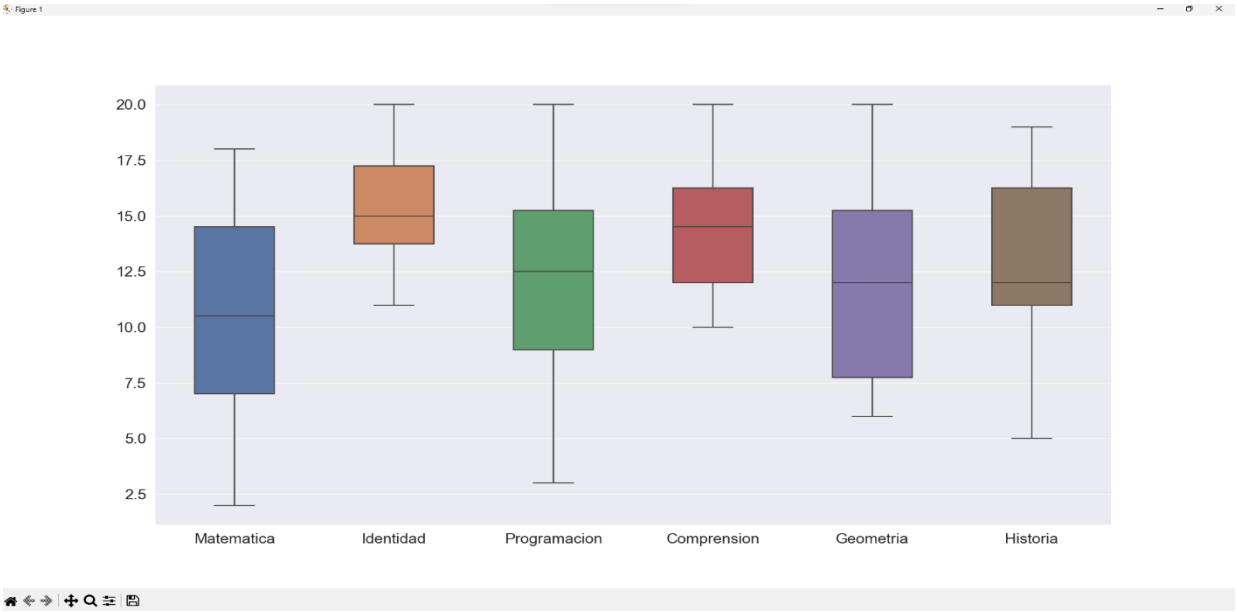

Figura 1. Gráfico de Caja y Bigote

Se puede observar en la figura 1, se puede observar que las materias con mayor dispersión de los datos son Geometría y Matemática, a su vez la materia cuyas notas tienen menor dispersión es identidad y le sigue Comprensión. Ambas materias son de orden Humanistíco.

Luego, se realizó un gráfico de Correlación con histograma para encontrar las correlaciones entre las diferentes notas que se obtuvieron para las asignaturas.

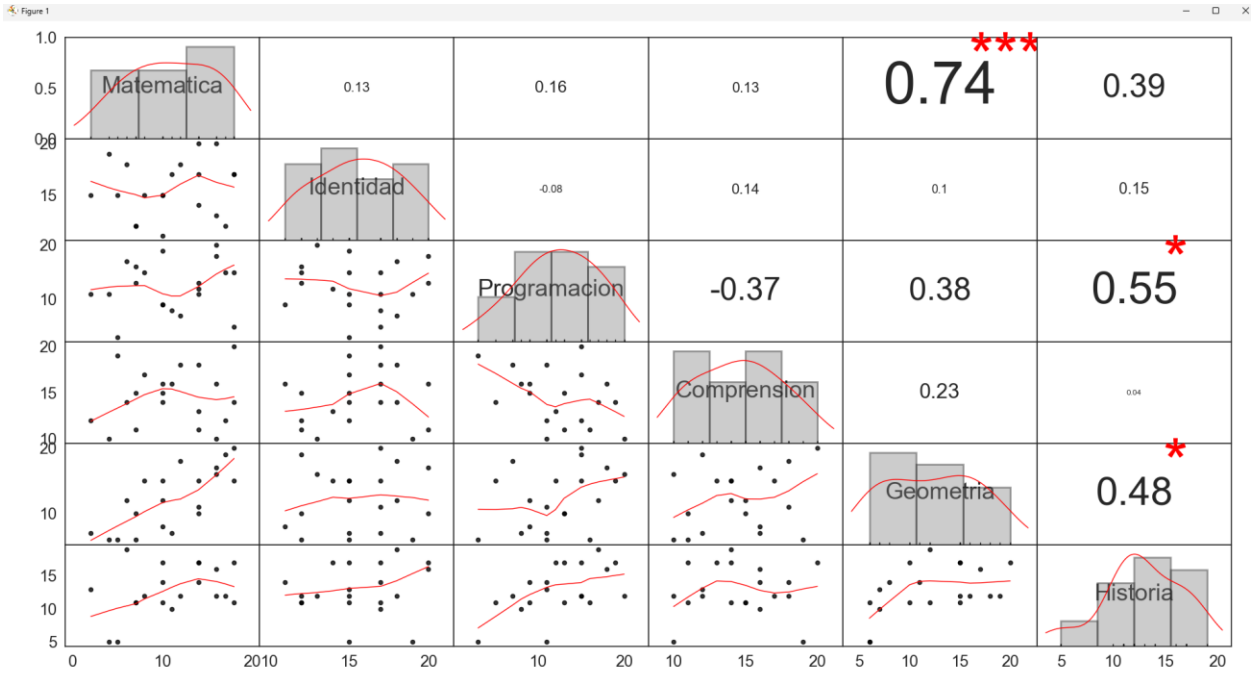

Figura 2. Gráfico de la Matriz de Correlación.

Se puede observar en el resultado de la figura 2, lo siguiente:

Analizando las mayores correlaciones positivas

- Matemática y Geometría tienen la mayor correlación positiva y es de 0.74, lo que da a entender que solo se necesita saber las notas de matemática para lograr predecir las notas de geometría y viceversa. Por el hecho de que la correlación es positiva también se puede decir que, si las notas de geometría son buenas, las de matemática también y viceversa, ya que estas son proporcionales. Programación e Historia tienen la segunda correlación más grande (0.55) dentro de la matriz de Correlaciones y por último Geometría e Historia tienen la tercera correlación más grande (0.48) dentro de la matriz de correlaciones.

Analizando las mayores correlaciones negativas

- Programación y Comprensión tienen la mayor correlación negativa (-0.37), lo que da a entender que solo se necesita saber las notas de programación para lograr predecir las notas de comprensión y viceversa. Por el hecho de que la correlación es negativa se puede decir que, si las notas de programación son altas, las de comprensión son bajas y viceversa, por último se tiene la correlación entre Identidad y programación (-0.08).

Basado en este análisis se puede concluir que solo se necesitan saber las notas de 2 materias que son Matemática y Programación para predecir las demás notas, sin embargo solo para el caso de matemática y

geometría se obtendrá una predicción muy acercada al verdadero resultado ya que su correlación es fuerte porque pasa de 0.7, para el resto de notas de las otras asignaturas se tendrá una idea no tan precisa de las notas ya que la correlación entre las demás notas de las asignaturas son débiles (por debajo de 0.7 o por encima de -0.7)

A continuación, se estandarizaron los datos para que todas las variables tuvieran la misma escala y se aplicó el PCA utilizando la clase PCA de sklearn. Se determinó el número de componentes principales a retener mediante el método de la varianza acumulada, que indica cuánta varianza se explica con cada componente adicional.

Se graficó la varianza acumulada para visualizar cómo la cantidad de varianza explicada aumenta a medida que se agregan más componentes principales. En este caso, se decidió conservar tres componentes principales, ya que explican aproximadamente el 81% de la varianza total. (Figura 3)

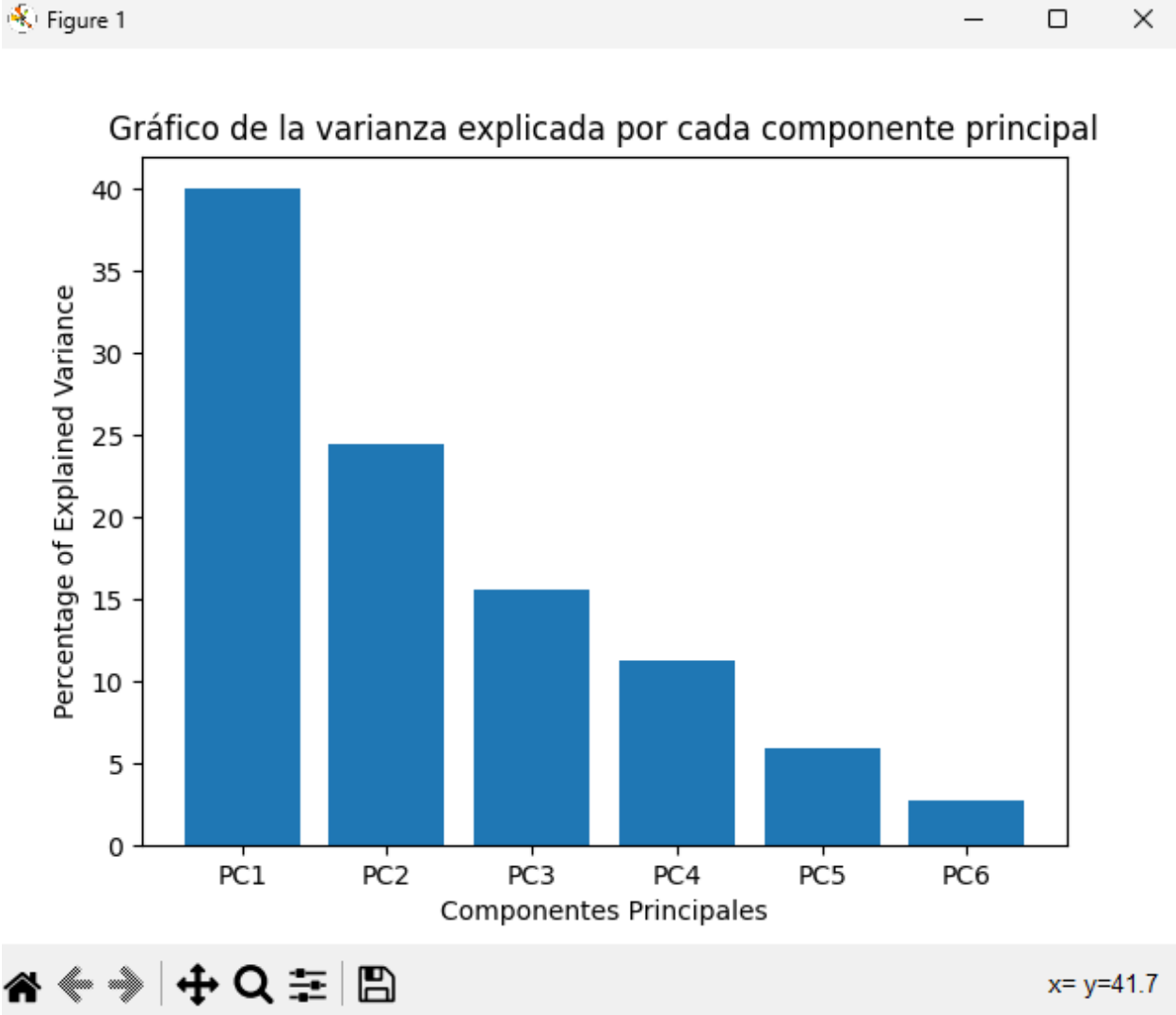

Figura 3. Gráfico de Componentes Principales

Finalmente, se graficó un gráfico de dispersión para visualizar la relación entre las tres primeras componentes principales. En este gráfico, se puede ver que las asignaturas Matemáticas, Programación y Geometría están fuertemente relacionadas entre sí y forman una agrupación clara. Las asignaturas de Identidad, Comprensión e Historia también están relacionadas entre sí, aunque no tan fuertemente como las anteriores. Además, se observa que los estudiantes que tienen altas calificaciones en la primera agrupación de asignaturas tienden a tener calificaciones bajas en la segunda, y viceversa. (Figura 4)

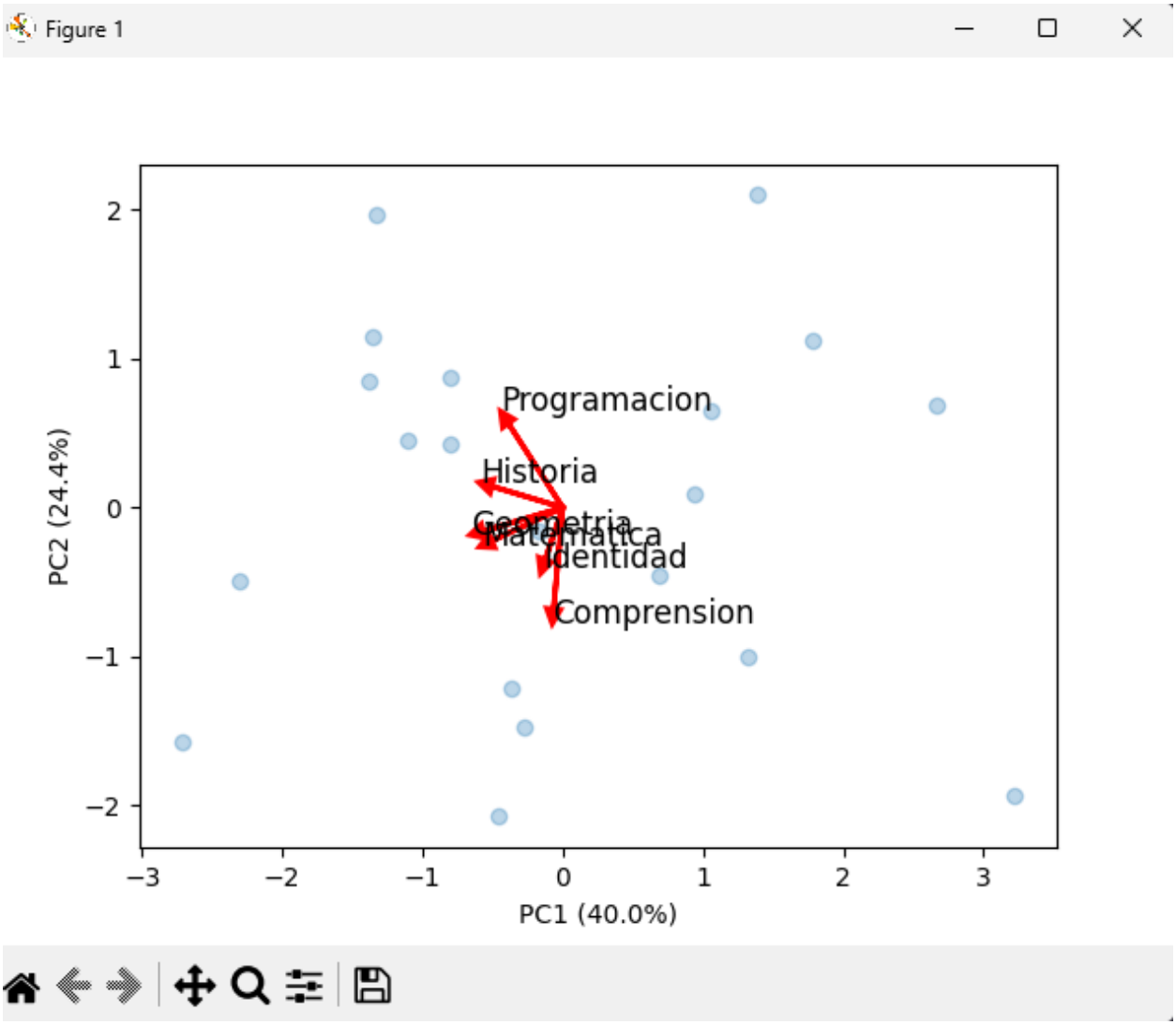

Figura 4. Gráfico de Dispersión

En conclusión, el análisis de componentes principales proporcionó una forma de resumir y visualizar la relación entre múltiples variables. Al retener tres componentes principales, se logró explicar la mayoría de la varianza total de los datos. La visualización de los datos en un gráfico de dispersión mostró que algunas asignaturas estaban fuertemente relacionadas entre sí y gracias al gráfico de la matriz de correlaciones se pudo determinar cuáles (Matemática y Geometría), mientras que otras estaban menos relacionadas (Comprensión e Historia). Además, se encontró una relación inversa entre las dos agrupaciones de

asignaturas, lo que sugiere que los estudiantes que tienen éxito en un conjunto de asignaturas pueden tener dificultades en otro conjunto. Dichas materias son (Programación y Comprensión).

Para Países:

En este análisis se explorará la estructura de un conjunto de datos que contiene información sobre la distribución de la agricultura, industria y servicios en 17 países de América. El objetivo de este análisis de componentes principales es encontrar patrones y relaciones ocultas entre las variables originales y reducir la dimensionalidad del conjunto de datos para facilitar su interpretación. Para ello, se utilizarán diversas técnicas como la matriz de correlación, la representación gráfica de los datos mediante diagramas de caja y bigote y gráficos de dispersión, y la aplicación de un análisis de componentes principales. El análisis de los resultados obtenidos permitirá determinar las variables que están más relacionadas entre sí y cómo se distribuyen en la muestra de países analizados.

Primero se realizó un diagrama de Caja y Bigote (Figura 5)

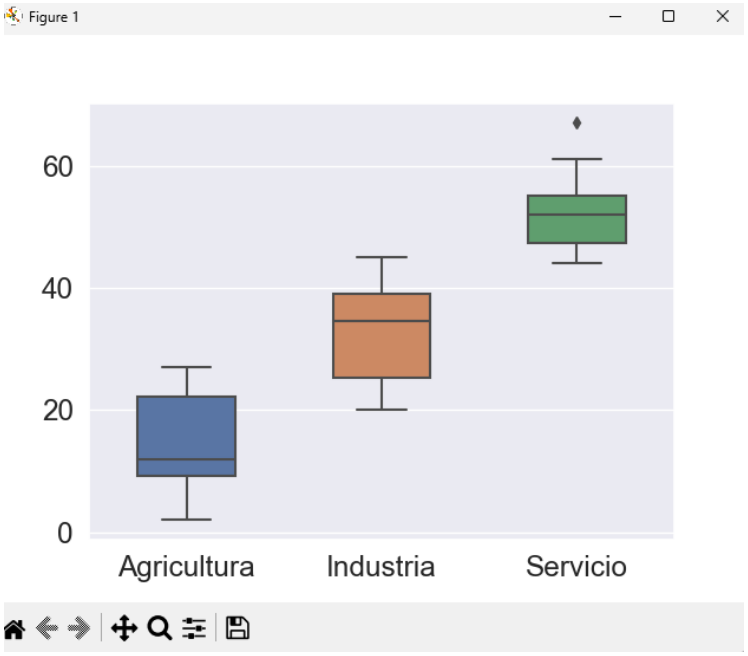

Figura 5. Grafico de caja y bigote

En el grafico se puede observar que, en agricultura como tiene la mayor dispersión quiere decir las inversiones no se parecen tanto entre los diferentes países e invierten menos del 15% en agricultura. En servicios debido a que su dispersión es la menor de todas quiere decir que las inversiones entre los países son mas parecidas y tienden a invertir mas en servicios que en agricultura e industria. El 50% de los países invierten menos del 35% en industria

Luego se graficó la matriz de correlación (Figura 6)

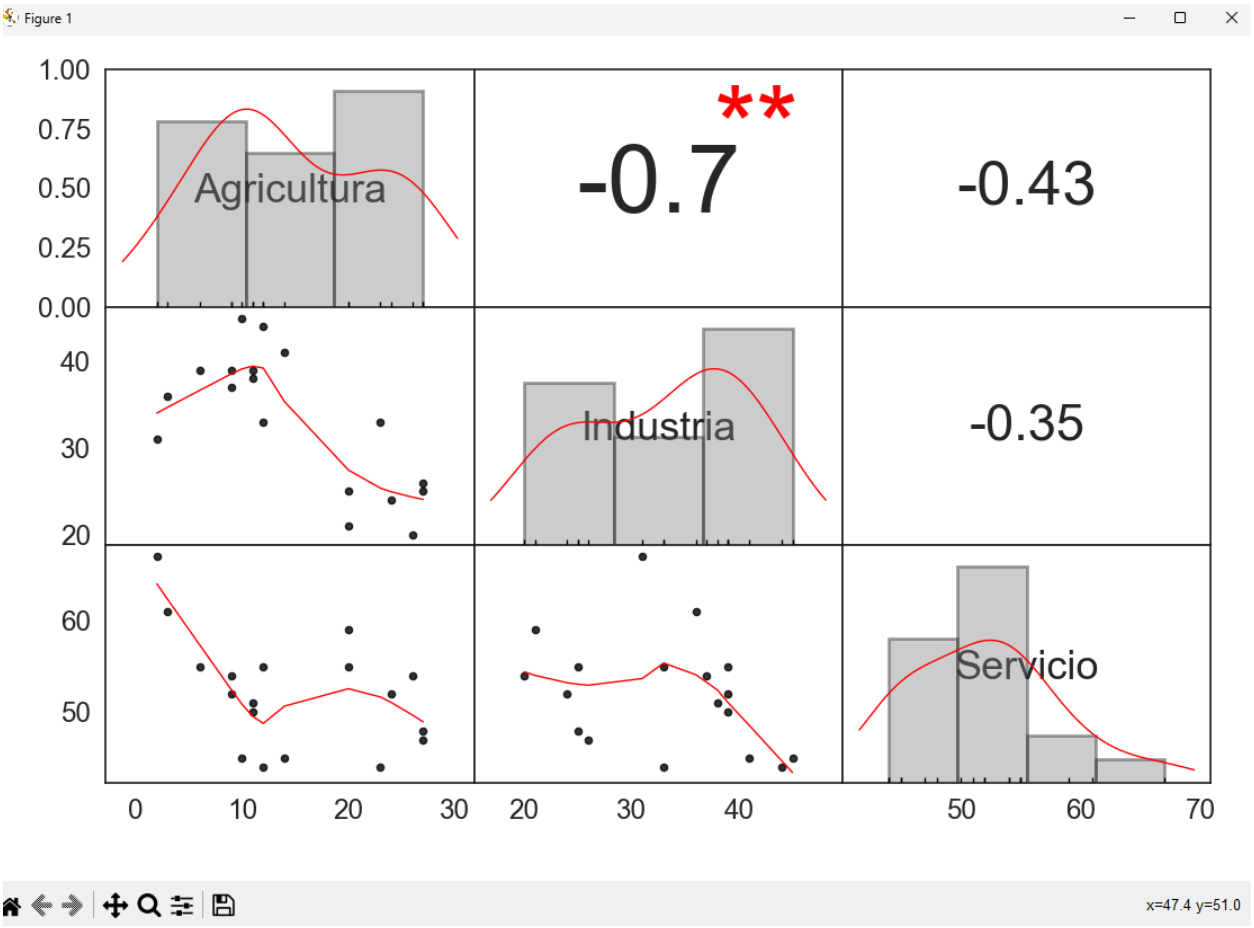

Figura 6. Matriz de Correlación

Analizando la figura se puede decir lo siguiente:

- Todas las correlaciones son negativas, lo que indica que cada uno de los PIB de los diferentes países son inversamente proporcionales, cuando uno baja el otro sube.
- Solo se necesitaba saber el valor de agricultura para predecir con una alta precisión el valor de industria y viceversa, debido a que su correlación es fuerte ya que es menor o igual a -0.7, a diferencia de industria y servicios que su correlación es mayor que -0.7, se puede predecir el valor, pero con una precisión mucho menor debido que es una correlación débil.

Posteriormente, se realizó el grafico de componentes principales (Figura 7)

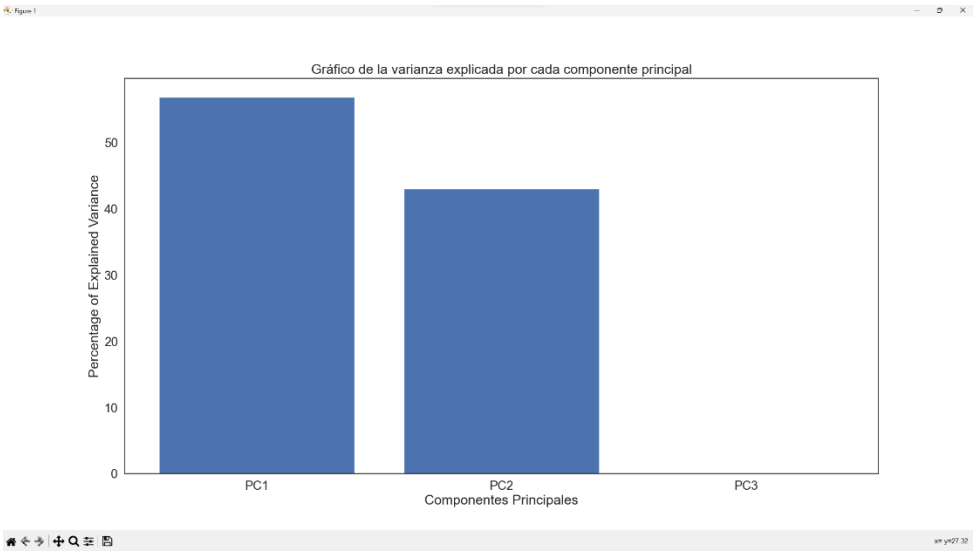

Figura 7. Gráfico de Componentes principales

Por ultimo se realizo el grafico de dispersión (Figura 8)

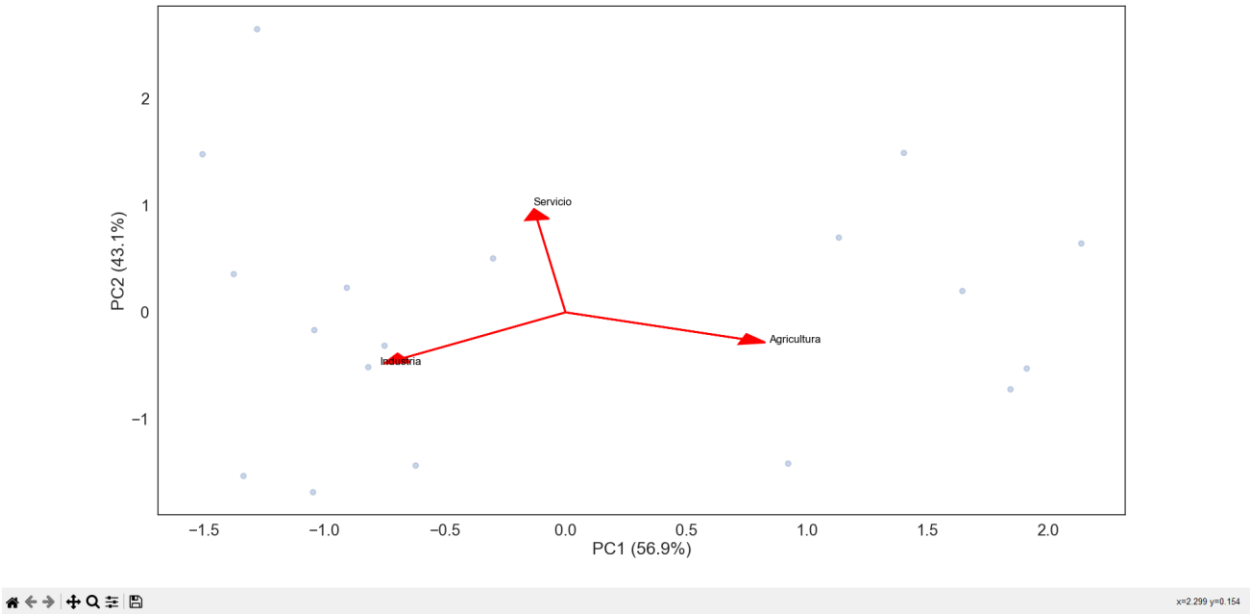

Figura 8. Gráfico de Dispersión

En el gráfico, se puede observar que existe una fuerte correlación negativa entre la agricultura y los servicios, y una correlación negativa moderada entre la agricultura y la industria. También se puede observar una correlación negativa moderada entre la industria y los servicios.

A partir de los gráficos y la matriz de correlación, se puede concluir que la variable de servicio es la más relevante para la explicación de los datos, seguida de la agricultura y la industria. También se puede

observar que la agricultura y los servicios tienen una fuerte correlación negativa, mientras que la industria no tiene una correlación muy fuerte con ninguna otra variable.

En conclusión, la economía de los países en este conjunto de datos parece estar dominada por el sector de servicios, con una correlación negativa fuerte con el sector de agricultura. La correlación entre la agricultura e industria es moderada y la industria no está muy correlacionada con ninguna otra variable.

Usando el conjunto de datos `europa`, realice un análisis descriptivo completo, un análisis de correlación y un análisis de componentes principales. Establezca conclusiones generales.

Análisis Descriptivo:
Para realizar el análisis descriptivo se necesita lo siguiente:

- Calcular la media, moda, mediana
- Calcular el rango intercuartílico
- Máximo
- Mínimo
- Rango
- Varianza
- Desviación estándar

Luego de calcular todos los valores anteriormente mencionados se obtuvo la siguiente tabla:

Variable	Media	Mediana	Moda	Rango intercuartílico	Máximo	Mínimo	Rango	Varianza	Desviación estándar
Agr	14.3	11.85	-	16.2	48.7	2.7	46	214.16	14.64
Min	1	0.8	-	1.1	3.1	0.1	3	0.41	0.64
Fab	27	27.6	-	4.9	41.2	16.8	24.4	37.85	6.15
Ene	0.9	0.9	-	0.5	1.4	0.5	0.9	0.06	0.25
Con	8.5	8.4	-	2.6	16.9	5.9	11	12.87	3.59
IS	13.6	14.4	-	5.6	19.1	6.1	13	32.45	5.7
Fin	4.6	5	-	1.5	8.5	0.5	8	8.81	2.97
SSP	19.1	20.1	-	6.3	32.4	4	28.4	35.71	5.98
TC	6.56	6.7	-	1.82	9.4	4	5.4	1.32	1.15

Es importante tener en cuenta que, para la variable Agr, hay una observación atípica con un valor de 48.7, lo que ha aumentado significativamente el rango, la varianza y la desviación estándar. Además, para la variable SSP, también hay una observación atípica con un valor de 32.4, lo que ha aumentado el rango, la varianza y la desviación estándar.

Análisis de Correlación

Primero se realizo el grafico de Caja y Bigote (Figura 9)

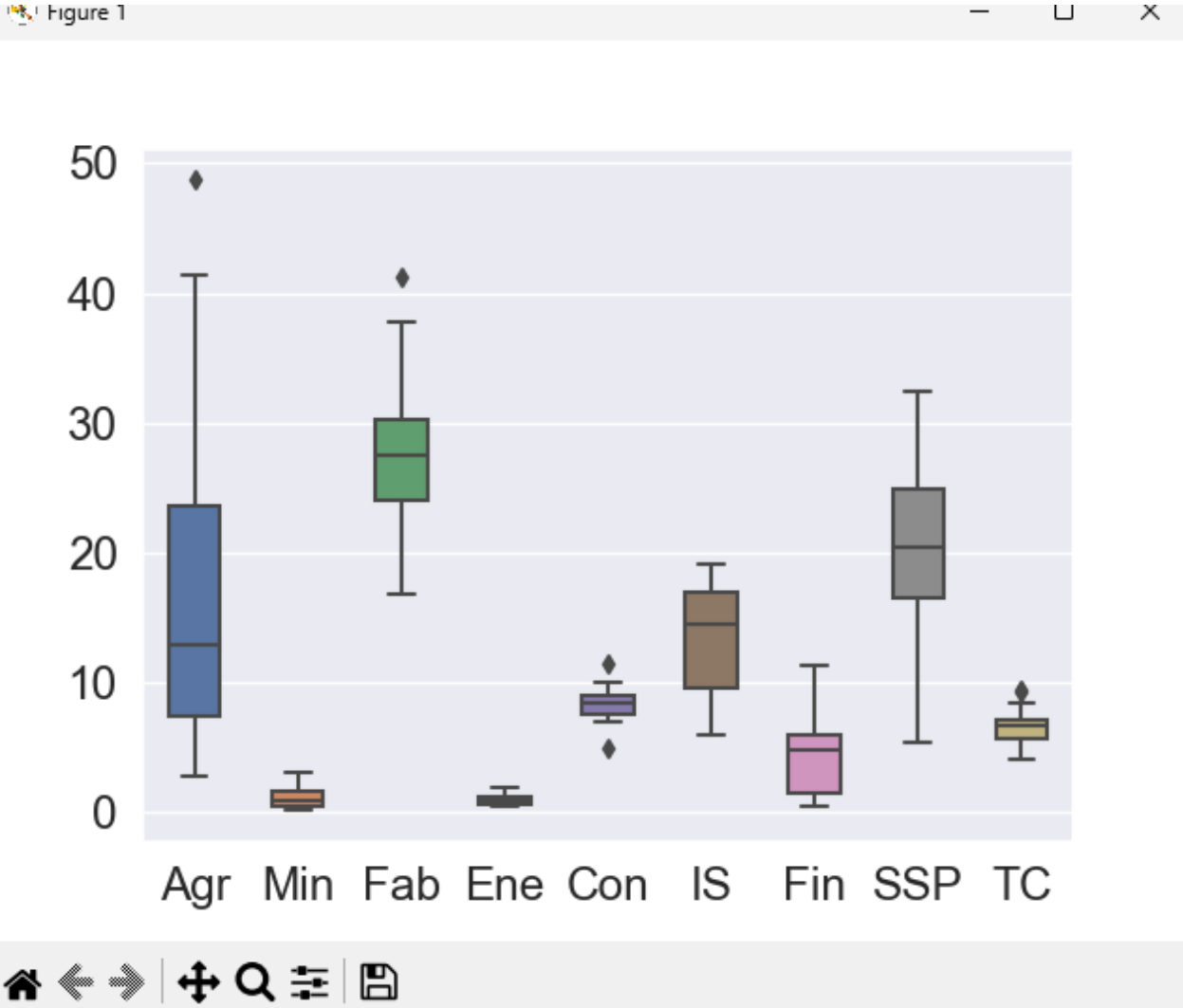

Figura 9. Grafico de Caja y bigote

Se puede observar que para `Ene` o `Energía`, la dispersión es la menor de todas lo que quiere decir que, la mayoría de los valores están muy cerca de la mediana y que hay pocos valores atípicos (outliers). Lo que indica que los datos son más homogéneos. Además, son los valores más pequeños estando por debajo del 10%, es decir, que los países invierten menos del 10% en Energía.
Para `Agr` o `Agricultura` los valores de esa variable están más dispersos, lo que implica una mayor variabilidad en los datos. Esto puede ser indicativo de una mayor variación o incertidumbre en los datos, lo que puede hacer que sea más difícil obtener conclusiones precisas o tomar decisiones basadas en esos datos. A su vez al tener los bigotes más largos da a entender que la variabilidad de los datos es alta y que hay una gran cantidad de valores atípicos o fuera de lo común en los datos.

Se realizo el grafico de Matriz de correlación (Figura 10)

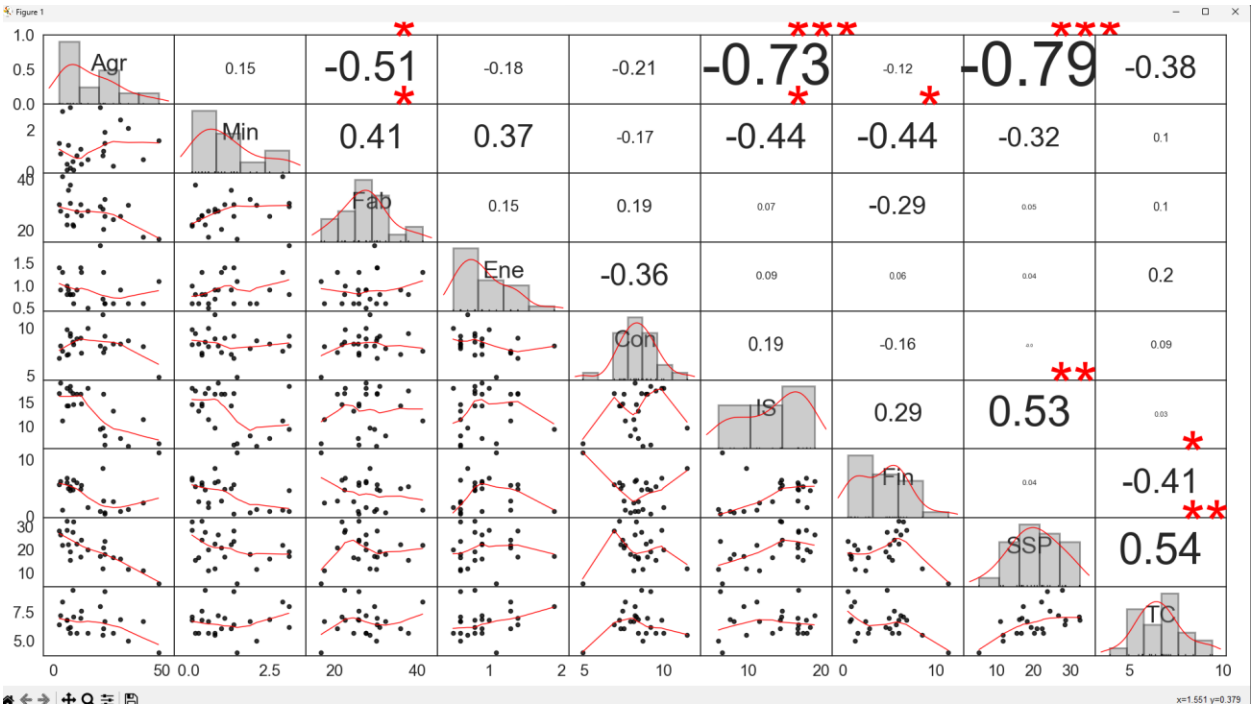

Figura 10. Gráfico matriz de Correlación

Se puede observar que hay 2 correlaciones fuertes, entre Agr e IS existe una correlación negativa de -0.73, lo que da entender que son inversamente proporcionales, y a su vez con solo saber el valor de uno de los dos se puede obtener el valor del otro, por ser una correlación negativa si el valor de Agr es alto el de IS es bajo y viceversa. Por otro lado, se tiene la correlación negativa mas alta que es entre SSP y Agr, y es el mismo caso expuesto anteriormente, cuando el valor de SSP sea bajo el de Agr será alto y viceversa, con el detalle que la predicción del otro valor será más preciso porque la correlación es mayor.
También se puede observar las correlaciones positivas, en las cuales resaltan las correlaciones entre SSP y TC, IS y SSP, lo que genera que se puede saber el valor de TC si se tiene el de SSP y viceversa, lo que lleva a concluir que se puede obtener el valor de IS y TC teniendo únicamente el valor de SSP, pero la predicción de dicho valor no será la mas precisa debido a que la correlación está por debajo de 0.7 lo que la clasifica como débil.

Análisis de Componentes principales:

Se realizo el grafico de Componentes principales varianza acumulada para visualizar cómo la cantidad de varianza explicada aumenta a medida que se agregan más componentes principales. En este caso, se decidió conservar cuatro componentes principales, ya que explican aproximadamente el 83% de la varianza total. (Figura 11)

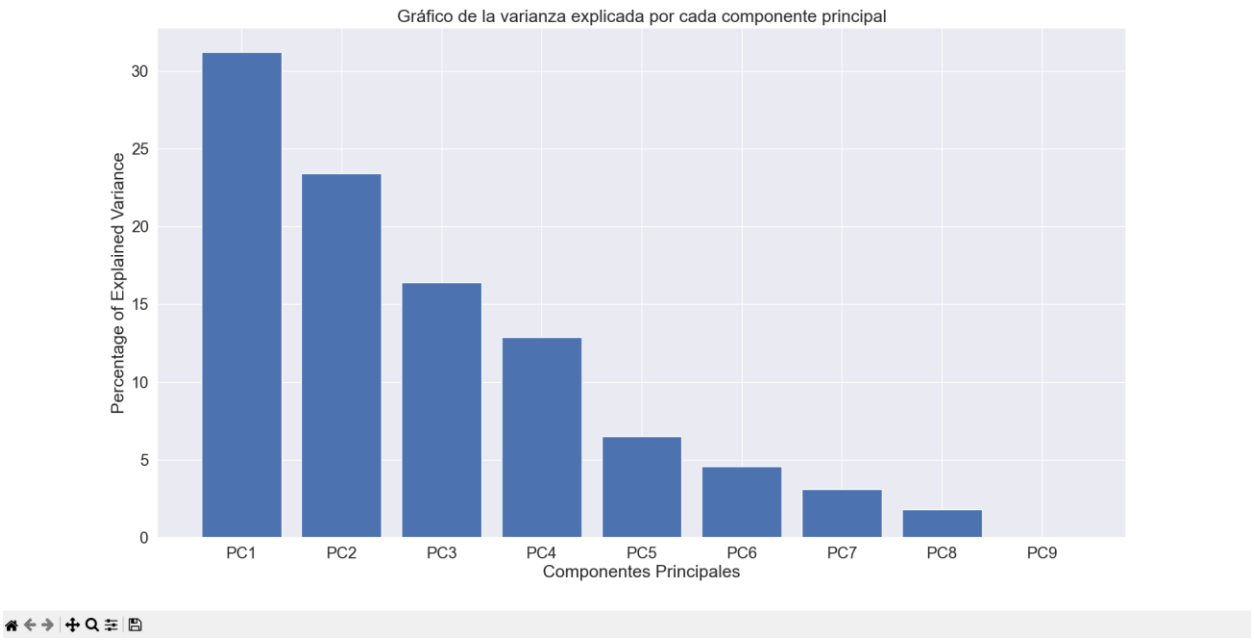

Figura 11. Gráfico de Componentes principales

Por ultimo se realizo el grafico de dispersión (Figura 12)

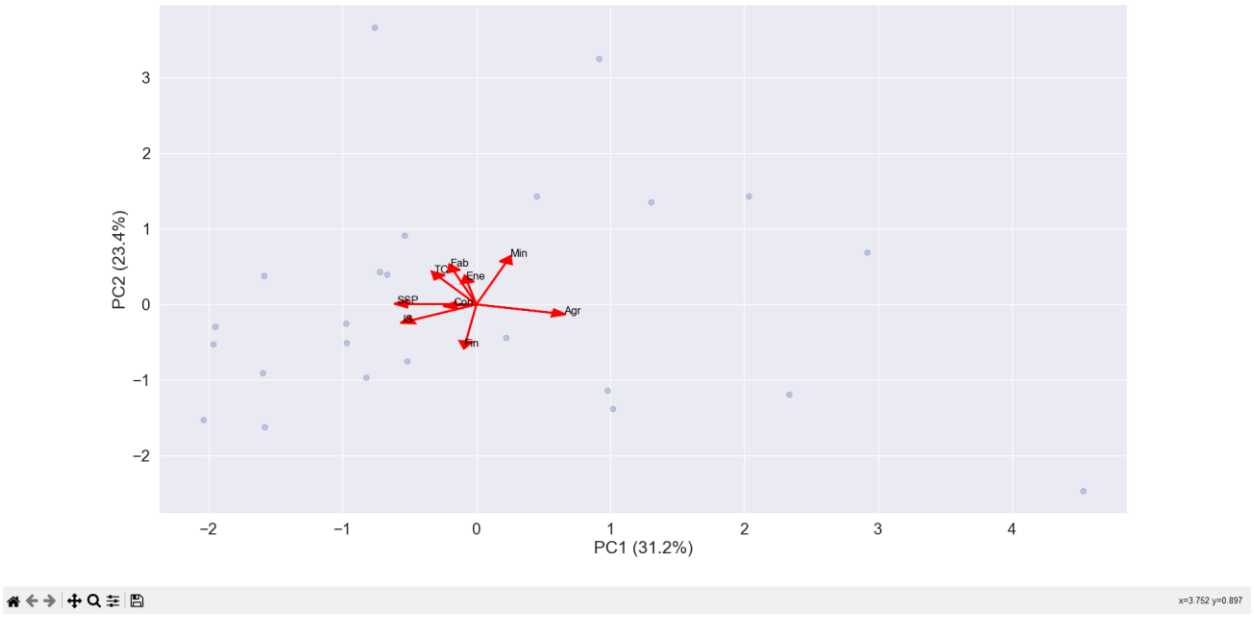

Figura 12. Gráfico de Dispersión

Al observar este grafico se puede notar que `Agr` tiene una fuerte correlación negativa con `SSP` y gracias a la matriz de correlaciones se sabe el valor exacto de la correlación (-0.79), además se puede ver que `SSP` y `TC` tienen una correlación positiva moderada y es de (0.54). También se puede observar que `IS` y `Agr` tienen una correlación negativa moderada.

En conclusión, se puede decir que los países son mas propensos a invertir en `Fab` que, en los otros servicios, teniendo unos valores similares entre todos los países debido a la dispersión de los datos, por otro lado, los países tienen a invertir menos en `Ene` pero con un conjunto de datos mucho más similar que en `Fab` nuevamente debido a la dispersión de dicha variable. `Agr` por tener la mayor cantidad de dispersión de todo el conjunto de datos se puede decir que mas del 12% aproximadamente de los países invierten, pero con valores muy diferentes causado por la dispersión que existe. También se puede concluir que cuando la inversión en `Agr` es alta, la inversión en `SSP`, `Fab` e `IS` son bajas debido a la alta correlación negativa que tienen, es decir son inversamente proporcionales. Lo que resulta en que si se sabe el valor de `Agr` se puede predecir el valor de `SSP`, `Fab` e `IS` con una precisión muy exacta para `SSP` e `IS` y no tan precisa para Fab. Por último, se puede concluir que Fab y Min al tener una correlación positiva, cuando la inversión de 1 de ellos es alta la del otro también, a su vez `IS`, `SSP` y `TC` tienen una correlación positiva, lo que logra permitir que si se conoce el valor de SSP se puede conocer el valor de `IS` y `TC` con una precisión medianamente acertada debido a que sus correlaciones están entre moderadas y débiles.

Referencias:

Bryner, R. J., & Ye, X. (2020). The impact of teaching quality on student academic achievement: Evidence from a US university. International Journal of Educational Research, 99, 101507.

McLeod, S., Weininger, E. B., & Martorell, F. (2020). Students' attitudes toward online education during the COVID-19 pandemic: A survey study of US undergraduate and graduate students. PloS one, 15(12), e0242900.

José A. Ruipérez-Valiente. (2020). El Proceso de Implementación de Analíticas de Aprendizaje. Enlace: https://www.redalyc.org/journal/3314/331463171005/html/

Gabriela Sabulsky. (2019). Universidad Nacional de Córdoba (UNC), Argentina. Analíticas de aprendizaje para mejorar la enseñanza y el seguimiento a través de entornos virtuales. Acerca de las analíticas de aprendizaje. Enlace: https://rieoei.org/RIE/article/download/3340/4029/

Castillo García, Manuel; Ramos Corpas, Manuel Jesús; Revuelta Marchena, Manuel. (2016). El análisis de datos educativo. Estrategias y técnicas para su realización como fase previa a la acción supervisora. Enlace: https://rio.upo.es/xmlui/handle/10433/4223