Descripción de los Datos

Para esta actividad se ah optado por utilizar un conjunto de datos con multiples variables, este consiste en observaciones de distintas variables económicas y laborales. Las variables incluidas son 'capital', 'labour', 'output' y 'wage', Cada una de estas es importante las industrias. Estas variables son todas numéricas y cuantitativas, lo que las hace adecuadas para el análisis multivariante

Preprocesamiento de los Datos

Limpieza y Estandarización

En primer lugar se analizo y estudio la base de datos buscando valores NAN, a lo que detectamos que estaba completa. Después de recoger todos los datos, decidimos darles un pequeño ajuste. Este ajuste, conocido como estandarización, es como el heroe no reconocido en el mundo del analisis de datos. Es especialmente importante cuando trabajamos con PCA, ya que nos ayuda a poner todas las variables en la misma línea de partida.

Análisis Exploratorio

Se realizaron analisis exploratorios iniciales, incluyendo la creación de diagramas de caja y de dispersion. Los diagramas de caja proporcionaron una visión general de la distribución de cada variable, teniendo en consideración valores anormales. Los diagramas de dispersión se utilizaron para observar las relaciones bivariadas entre las variables.

Análisis Multivariante

PCA (Análisis de Componentes Principales)

Usamos PCA para simplificar nuestro conjunto de datos y encontrar los componentes principales que explican la mayor parte de la variabilidad en los datos. Luego, estudiamos cuánta diferencia encontrábamos con el componente principal, gracias a esto fuimos capaces de decidir cuántos componentes mantener. Visualizamos esto con un gráfico llamado screeplot. Finalmente, miramos los biplots para entender las relaciones entre las variables y los componentes, y para ver cómo se agrupaban las diferentes observaciones en el espacio de los componentes principales.

**Interpretación de Resultados**

Discusión de los Hallazgos

* **Vectores Medios y Distribución de Datos**: Los vectores medios para 'capital', 'labour', 'output' y 'wage' muestran la tendencia central de cada variable. Con estos valores observamos un promedio de las variables y nos ayuda a entender el contexto.
* **Matriz de Datos Centrada**: esta es utilizada para ver como se comporta la desviación de cada media y nos ayuda a luego poder hacer un análisis PCA.
* **La matriz de varianza-covarianza nos muestra que hay bastante variabilidad y relaciones lineales entre las variables. En cambio con la matriz de correlación podemos ver que hay fuertes correlaciones positivas entre ciertas variables, como ‘capital’, ‘labour’ y ‘output’. Esto significa que estas variables suelen moverse juntas.**
* **Estandarización de Variables**: La estandarización de las variables asegura que cada una contribuya equitativamente al análisis. La transformación resultante muestra cómo las observaciones se distribuyen alrededor de una media de cero y una desviación estándar de uno.
* **Distancias Euclidianas**: La matriz de distancias euclidianas proporciona una medida de la disimilitud entre las observaciones. Las distancias más cortas indican observaciones más similares, mientras que las distancias más largas sugieren mayor divergencia entre los puntos de datos.
* **Descomposición de Valores Singulares (SVD)**: La aproximación de rango 3 a la matriz de datos centrada, obtenida a través de SVD, ofrece una visión simplificada de los datos, capturando los aspectos más críticos de la variabilidad en tres dimensiones.
* **Variable Dicotómica**: La creación de una variable dicotómica basada en la mediana de 'labour' añade una dimensión categórica al análisis, permitiendo examinar cómo se distribuyen los datos en torno a este punto de corte.
* **PCA y Loadings**: El PCA reveló que los primeros componentes capturan una proporción significativa de la varianza en los datos. Los 'loadings' de los componentes principales muestran la contribución de cada variable a estos componentes, donde 'output' tiene una fuerte influencia en el primer componente.
* **Varianza de las Scores**: La varianza de las 'scores' extraídas indica la cantidad de información capturada por cada componente principal, con el primer componente dominando en términos de explicar la variabilidad en los datos.
* **Correlación entre Variables y Componentes Principales**: La matriz de correlación entre las variables originales y los componentes principales muestra cómo cada variable original se relaciona con los componentes, proporcionando una visión de cómo las dimensiones originales se transforman en el espacio del componente principal.

**Conclusión**

Resumen de los Hallazgos

Los hallazgos indican fuertes relaciones lineales entre las variables económicas y laborales en el conjunto de datos. El PCA identificó componentes principales que capturan eficazmente la varianza en los datos, con 'output' mostrando una influencia significativa.

Conclusiones Generales

Este análisis proporciona una visión valiosa de las interacciones entre las variables estudiadas. Futuras investigaciones podrían explorar la aplicabilidad de estos hallazgos en diferentes contextos económicos o utilizar otros métodos de análisis para comparar y contrastar los











