Julián Andres Eduardo Arana Guiza
Anamaria Leguizamon Alarcon
Daniel Andrés Becerra Sierra
Juan David Lopez Becerra

1. **Comprensión del negocio:**

   Latinoamérica, conformada por más de 20 países, es una región sumamente diversa, tanto en términos culturales como socioeconómicos, características demográficas y económicas. En la última década, ha experimentado cambios significativos que han afectado el crecimiento urbano, la calidad de vida y la desigualdad. Para realizar un análisis completo de las ciudades latinoamericanas, se han seleccionado una serie de variables clave que permiten comprender su dinámica en términos de demografía y urbanización, salud y desarrollo humano, economía y empleo, desigualdad y pobreza, y medio ambiente.

   En el ámbito demográfico, la densidad de población, la tasa de crecimiento poblacional y la superficie total de la ciudad son fundamentales para entender la estructura y distribución de las personas en el espacio urbano. Estas variables nos ayudan a identificar patrones de crecimiento y a evaluar el grado de presión que la población ejerce sobre los servicios y recursos disponibles.

   En cuanto a salud y desarrollo humano, variables como la esperanza de vida y la tasa de alfabetización son indicadores directos del bienestar de la población. Una mayor esperanza de vida suele estar asociada con mejores servicios de salud y condiciones de vida, mientras que una alta tasa de alfabetización refleja el acceso a la educación y, en general, un mejor desarrollo humano.

   Para evaluar el ámbito económico, se consideran variables como el PIB, la tasa de desempleo, el costo de vida y el salario promedio. El PIB nos da una visión general de la actividad económica de la ciudad, mientras que la tasa de desempleo y el salario promedio proporcionan información sobre el mercado laboral y la calidad de los empleos. El costo de vida, por su parte, es esencial para entender el nivel de acceso que tienen los ciudadanos a bienes y servicios.

   La desigualdad y la pobreza son temas críticos en la región, por lo que es importante incluir indicadores que nos permitan medir estos fenómenos de manera precisa. Variables como la distribución del ingreso y los índices de pobreza nos muestran las disparidades económicas dentro de las ciudades y ayudan a identificar áreas que requieren políticas de intervención.

   Por último, el medio ambiente y el clima son factores que también influyen en la calidad de vida urbana. Variables como el índice de calidad del aire, la temperatura, la precipitación y la contaminación son esenciales para evaluar el impacto ambiental en la salud y el bienestar de la población. Además, la altitud puede tener un efecto significativo en el clima local y, por ende, en las condiciones de vida.

   Objetivos:

   Representación Estructurada:

   1. Desarrollar un perfil detallado de las principales ciudades de Latinoamérica basado en indicadores socioeconómicos, demográficos y ambientales.
   2. Identificar patrones que permitan categorizar las ciudades de acuerdo a su nivel de desarrollo y calidad de vida.
   3. Aplicar técnicas de reducción de dimensiones para identificar los principales factores que caracterizan a las ciudades latinoamericanas.

   Representación No Estructurada:

   1. Analizar las canciones más populares en español de la década 2010-2019 para extraer temas recurrentes y visualizar las relaciones semánticas entre las canciones.

2. Comprender las tendencias culturales y cambios en los temas musicales predominantes a lo largo de la década.
3. Utilizar técnicas de minería de texto para extraer e interpretar los temas más comunes en la música popular de la región.

Para llevar a cabo de manera exitosa nuestro proyecto, hemos dividido nuestro proyecto en 5 fases:

1. Recolección y limpieza de datos.
2. Análisis exploratorio de datos.
3. Aplicación de SVD para datos estructurados y técnicas de minería de texto para datos no estructurados.
4. Interpretación y visualización de resultados.
5. Documentación y preparación de la presentación final.

2. Comprensión de los datos:

Estructura de la tabla "Ciudad":

Diccionario de datos:

Campo	Tipo de Dato	Longitud	Descripción	Restricciones
Ciudad	VARCHAR	100	Nombre de la ciudad	NOT NULL, UNIQUE
Población	INT	N/A	Número total de habitantes	NOT NULL
Densidad de la población	DECIMAL	10,2	Número de habitantes por KM2	NOT NULL
Tasa de crecimiento Poblacional	DECIMAL	5,2	Tasa anual de crecimiento de la población (%)	NOT NULL
Esperanza de vida (promedio años)	DECIMAL	5,2	Esperanza de vida promedio en años	NOT NULL
Tasa de alfabetización (%)	DECIMAL	5,2	Porcentaje de la población alfabetizada	NOT NULL
PIB	DECIMAL	15,2	Producto Interno Bruto en USD	NOT NULL
Tasa de desempleo (%)	DECIMAL	5,2	Porcentaje de la población desempleada	NOT NULL
Salario promedio (USD)	DECIMAL	10,2	Salario promedio en USD	NOT NULL

Tasa de pobreza (%)	DECIMAL	5,2	Porcentaje de la población bajo la línea de pobreza	NOT NULL
Índice de calidad del aire (ICA)	DECIMAL	5,2	Calidad del aire, ICA	NOT NULL
Temperatura media anual (c°)	DECIMAL	5,2	Promedio de temperatura anual en grados Celsius	NOT NULL
Precipitación media nual (milímetros/año)	DECIMAL	10,2	Promedio anual de precipitación en mm	NOT NULL
Índice de contaminación	DECIMAL	5,2	Índice de contaminación general	NOT NULL
Escala Exp. de Contaminación	VARCHAR	50	Descripción cuantitativa exponencial de la contaminación	NOT NULL
Número de universidades	INT	N/A	Número total de universidades en la ciudad	NOT NULL
Tasa de criminalidad (crímenes por cada 100000 habitantes)	DECIMAL	5,2	Crímenes por cada 100000 habitantes	NOT NULL
Tasa de homicidios (homicidios por cada 100000)	DECIMAL	5,2	Homicidios por cada 100000 habitantes	NOT NULL
Superficie total (KM2)	DECIMAL	10,2	Superficie total de la ciudad en KM2	NOT NULL
Altitud Media (msnm)	DECIMAL	10,2	Altitud media sobre el nivel del mar (msnm)	NOT NULL
UTC Offset	VARCHAR	10	Diferencia horaria UTC	NOT NULL

Representación estructurada:

Preparación de los datos

Se parte de un conjunto de datos que contiene información sobre las ciudades, pero dado que la técnica de componentes principales sólo trabaja con variables numéricas, es necesario remover la columna con el nombre de las ciudades, así mismo se identifica si hay nulos que en este caso no se da, aun así se debe tener en cuenta que las escalas de los datos son muy diferentes por lo que al realizar el ACP es aconsejable emplear la matriz de correlación que ejemplifica una estandarización de las variables.

Análisis de la viabilidad para aplicar ACP

Antes de proceder, verificamos la viabilidad de utilizar ACP mediante la prueba de esfericidad de Bartlett. Esta prueba evalúa si las variables están correlacionadas, condición necesaria para aplicar esta técnica.

En este caso queremos probar que:

$$H_0 : \Sigma = I$$
$$H_1 : \Sigma \neq I$$

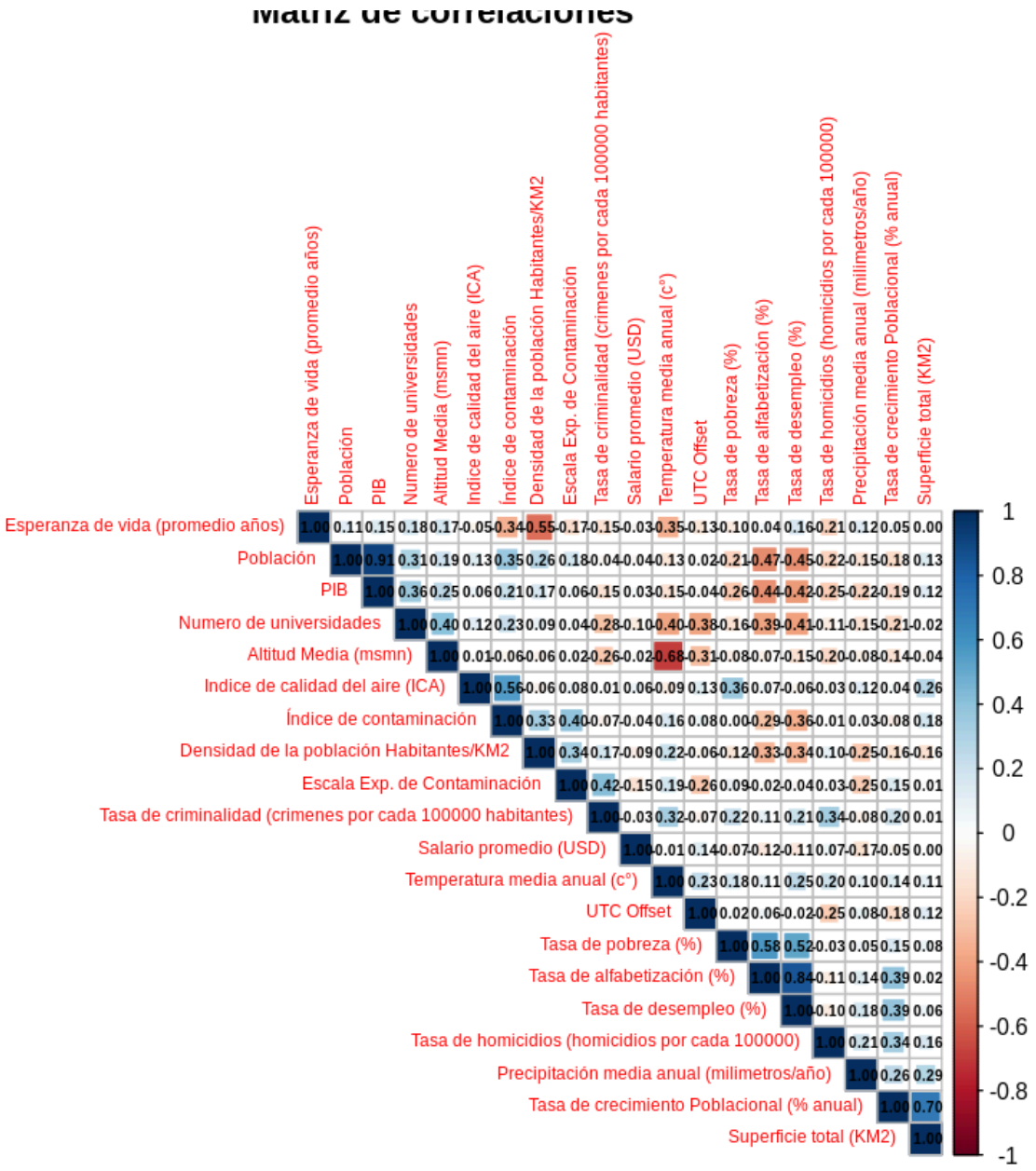
Y mediante la prueba tenemos los siguientes valores

```
$chisq
        502.78093298952
$p.value
        4.17519706806139e-30
$df
        190
```

Note que:

$$p - valor = 4.17519706806139e - 30 < 0.05 = \alpha$$

Por lo que podemos decir que hay suficiente evidencia en la muestra para rechazar la hipótesis nula, dandonos via para realizar el ACP
Si se gusta podemos ver las diferentes correlaciones que hay en la siguiente matriz de correlación:



Modelado:

Realizando el analisi de componentes principales mediante la función prcomp nos da los siguientes valores:

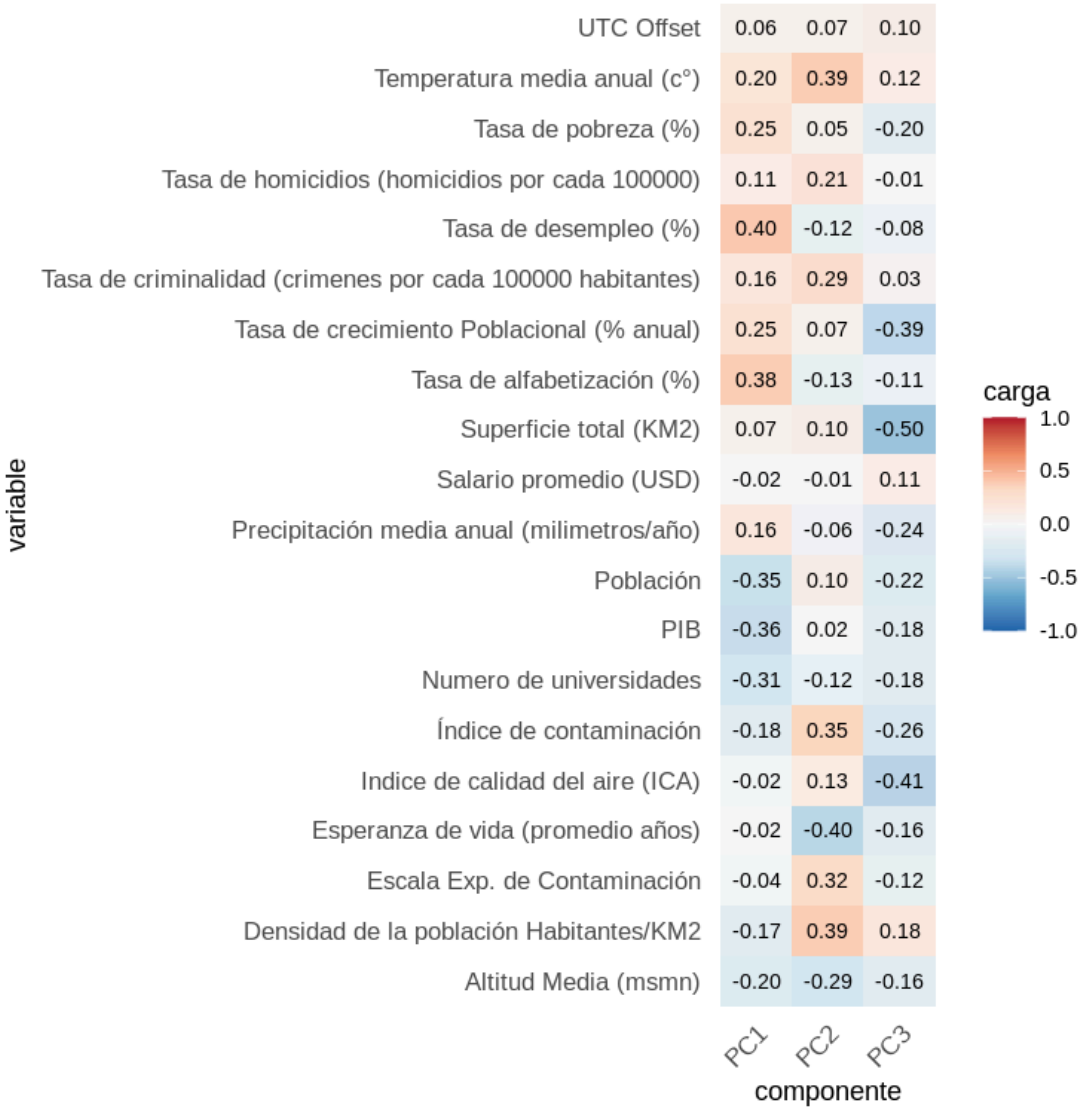
```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.0334 1.6709 1.4568 1.35166 1.30837 1.16995 1.06106
Proportion of Variance 0.2067 0.1396 0.1061 0.09135 0.08559 0.06844 0.05629
Cumulative Proportion  0.2067 0.3463 0.4525 0.54381 0.62940 0.69784 0.75413
                           PC8    PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.94943 0.8603 0.83872 0.75930 0.65049 0.60815 0.54751
Proportion of Variance 0.04507 0.0370 0.03517 0.02883 0.02116 0.01849 0.01499
Cumulative Proportion  0.79920 0.8362 0.87137 0.90020 0.92136 0.93985 0.95484
                          PC15    PC16    PC17    PC18    PC19    PC20
Standard deviation     0.53166 0.48566 0.40627 0.32672 0.25538 0.21839
Proportion of Variance 0.01413 0.01179 0.00825 0.00534 0.00326 0.00238
Cumulative Proportion  0.96897 0.98076 0.98902 0.99435 0.99762 1.00000
```

El análisis de componentes principales muestra que los primeros 10 componentes explican más del 90% de la variabilidad total, con el primer componente (PC1) capturando el 20.67% y los primeros 3 componentes explicando conjuntamente el 45.24%. La desviación estándar del PC1 es 2.0343, lo que implica que captura la variabilidad equivalente a aproximadamente 4.14 variables. A medida que avanzamos hacia componentes posteriores, la proporción de varianza explicada disminuye, con componentes como PC8 en adelante explicando menos del 5% cada uno y con desviaciones estándar cada vez menores. Esto sugiere que la mayor parte de la variabilidad está concentrada en los primeros componentes, siendo los más relevantes para capturar la estructura de los datos, y que utilizar más de 10 componentes probablemente no aporte mucha información adicional significativa.

Ahora empleando el criterio de la varianza acumulada, mediante un umbral regido por la ley de los pocos vitales, es decir del 80% hemos decidido quedarnos con **los primeros 9 componentes** los cuales atrapan una proporción de varianza explicada acumulada del **83%**

Evaluación

Podemos ver las cargas de las componentes en los siguientes graficos



PC1:
- Cargas positivas altas: Tasa de desempleo, Tasa de alfabetización, Tasa de crecimiento Poblacional y tasa de pobreza
- Cargas negativas altas: PIB, Población, Número de universidades

Este componente parece contrastar indicadores de desarrollo económico y educativo con crecimiento poblacional, pobreza y desempleo. Podríamos llamarla **"Mal desarrollo Socioeconómico".**

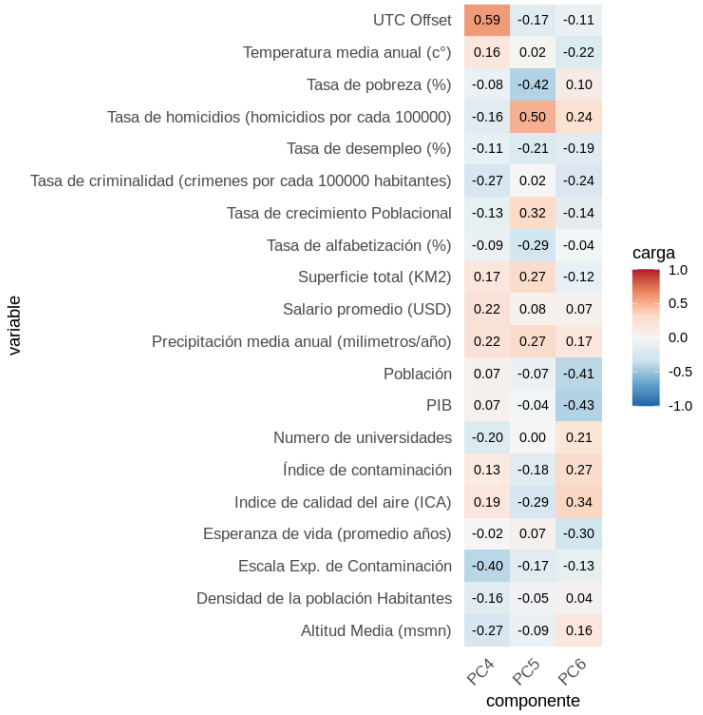
PC2:
- Cargas positivas altas: Temperatura media anual, Tasa de criminalidad, Índice de contaminación, escala exp de contaminación, Densidad de la población

- Cargas negativas altas: Esperanza de vida

Este componente parece relacionar factores urbanos como criminalidad, contaminación y densidad poblacional con clima cálido. Podríamos nombrarla, además de tener una relación interesante negativa con la esperanza de vida, como si en esta muestra por lo menos (aunque es lógico) todos estos factores como la criminalidad, contaminación y cantidad de habitantes tuvieran una relación inversa con la esperanza de vida, en este caso el nombre sería **"Características de las ciudades con menor esperanza de vida".**

PC3:
- Cargas positivas altas: casi todas son moderadas
- Cargas negativas altas: Superficie total, Índice de calidad del aire, Tasa de crecimiento Poblacional

Esta componente representa las ciudades al parecer más pequeñas Y con una calidad del aire buena, el nombre podría ser: **“ciudades pequeñas”**



PC4:
- Carga positiva alta: UTC Offset
- Cargas negativas altas: Escala Exp. de Contaminación , Tasa de criminalidad , Altitud Media

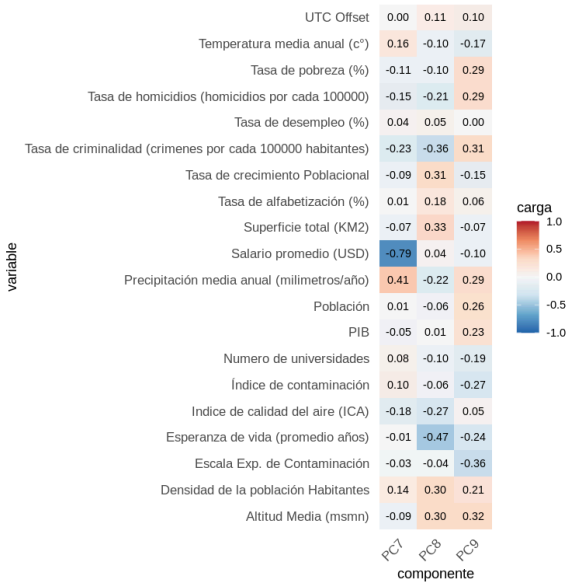
Este componente parece relacionar la ubicación geográfica (posiblemente más al este) con menor contaminación, criminalidad y altitud. Podríamos llamarla **"Localización Geográfica y Calidad Ambiental".**

PC5:
- Cargas positivas altas: Tasa de homicidios, Tasa de crecimiento Poblacional
- Cargas negativas altas: Tasa de pobreza , Tasa de alfabetización , Índice de calidad del aire

Este componente contrasta violencia y crecimiento poblacional con indicadores de desarrollo social y ambiental. Pareciera que a mayor tasa de homicidios y mayor población la pobreza es menor así como la alfabetización más sin embargo la calidad del aire mejora. Podríamos nombrarla **"Seguridad y Desarrollo Socioeconómico".**

PC6:
- Cargas positivas altas: Índice de calidad del aire , Índice de contaminación
- Cargas negativas altas: PIB , Población , Esperanza de vida

Esta componente parece relacionar menor calidad del aire y más universidades con menor PIB y población. Podríamos llamarla **"Contaminación contra indices socioeconomicos".**



PC7:
- Cargas positivas altas: Precipitación media anual
- Cargas negativas altas: Salario promedio

Este componente parece asociar mayor precipitación con menores salarios. Podríamos llamarla **"Ciudades con alta precipitación e ingresos bajos"**.

PC8:
- Cargas positivas altas: Tasa de crecimiento poblacional, Superficie total, Altitud media
- Cargas negativas altas: Tasa de criminalidad, Índice de calidad del aire, Esperanza de vida

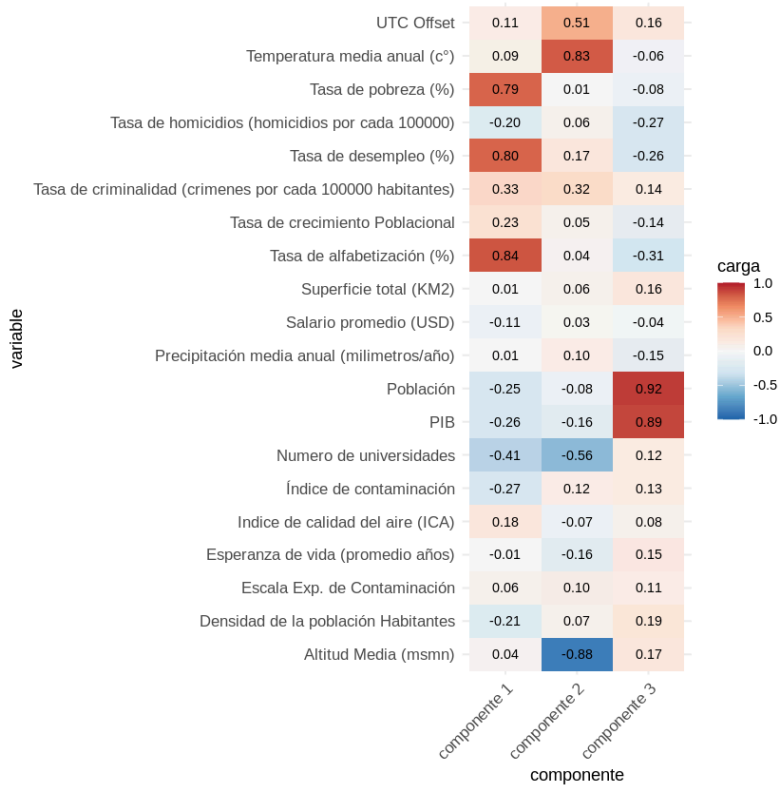
Este componente parece asociar ciudades con mayor crecimiento poblacional, superficie y altitud con menores tasas de criminalidad, mejor calidad del aire y mayor esperanza de vida. Podríamos llamarla **"Crecimiento y extensión contra calidad de vida"**.

PC9:
- Cargas positivas altas: Tasa de criminalidad, Tasa de pobreza, Precipitación media anual y altitud media

- Cargas negativas altas: Escala Exp. de Contaminación, Índice de contaminación, Esperanza de vida

Este componente parece asociar mayores tasas de criminalidad, pobreza y precipitación con menores niveles de contaminación y mayor esperanza de vida. Podríamos llamarla **"Problemas sociales por la calidad ambiental".**

Ahora para un mejor análisis emplearemos la interpretación de la rotación de los componentes



Componente 1:
- Cargas positivas altas: Tasa de pobreza, Tasa de desempleo, Tasa de alfabetización
- Cargas negativas altas: Índice de contaminación, PIB, Número de universidades

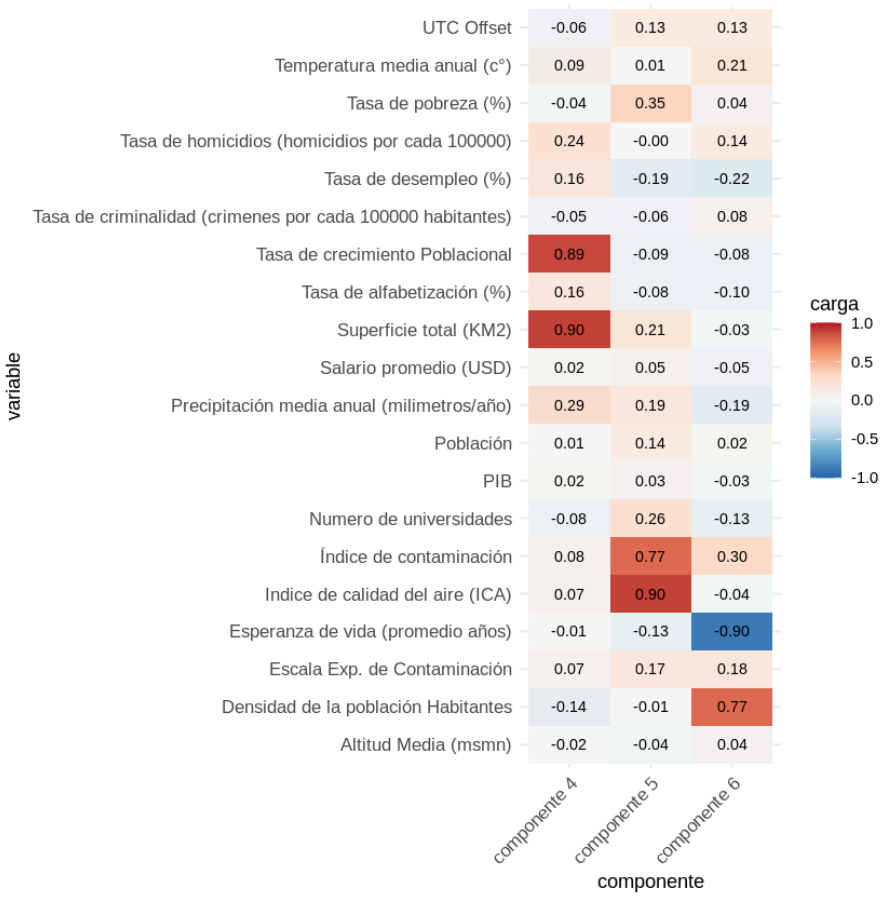
Este componente parece relacionar tasas sociales como pobreza, desempleo y alfabetización con una menor contaminación, PIB y número de universidades. Podríamos llamarla **"Desafíos socioeconómicos contra desarrollo urbano y educativo"**.

Componente 2:
- Cargas positivas altas: Temperatura media anual, UTC Offset
- Cargas negativas altas: Altitud media, Número de universidades

Este componente parece asociar ciudades con una mayor temperatura y una menor altitud, menor número de universidades, además de horarios cercanos al UTC 0. Podríamos llamarla **"Ciudades costeras"**.

Componente 3:
- Cargas positivas altas: Población, PIB
- Cargas negativas altas: Tasa de homicidios, Tasa de alfabetización

Este componente parece asociar una mayor población y PIB con una menor tasa de homicidios y alfabetización. Podríamos llamarla **"Ciudades con amplia población y PIB pero con baja alfabetización"**.



Componente 4:
- Cargas positivas altas: Tasa de crecimiento poblacional, Superficie total.
- Cargas negativas altas: Ninguna sobresaliente

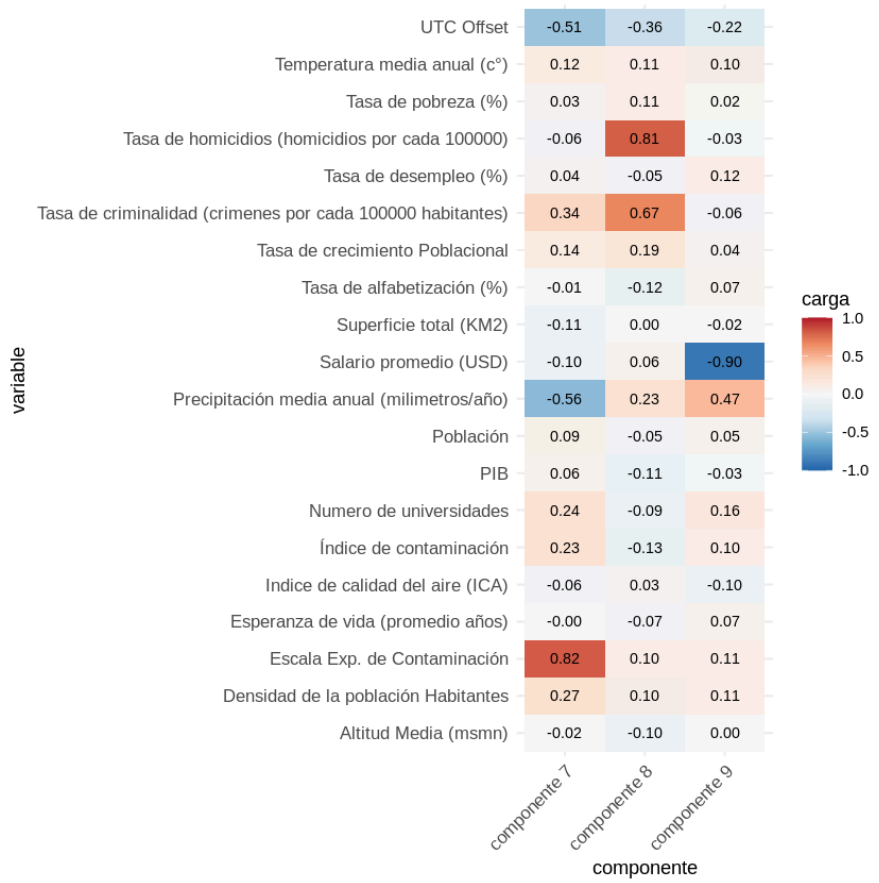
Este componente parece mostrar ciudades un mayor crecimiento poblacional, mayor superficie. Podríamos llamarla **"Crecimiento y expansión territorial de ciudades latinoamericanas"**.

Componente 5:
- Cargas positivas altas: Índice de contaminación, Índice de calidad del aire
- Cargas negativas altas: Ninguna sobresaliente

Este componente representa las ciudades con mayor contaminación aunque paradójicamente mejor calidad del aire, por lo que podríamos tomarla como un error de muestreo o no. Podríamos llamarla **"Ciudades con mayor contaminación pero mejor calidad del aire"**.

Componente 6:
- Cargas positivas altas: Densidad de la población
- Cargas negativas altas: Esperanza de vida

Este componente parece asociar una menor esperanza de vida con una mayor densidad de población, lo cual puede ser una característica interesante de la muestra. Podríamos llamarla **"Ciudades con mayor población pero menor esperanza de vida"**.



Componente 7:
- Cargas positivas altas: Escala Exp. de Contaminación
- Cargas negativas altas: Precipitación media anual, UTC Offset

Este componente parece asociar una mayor tasa de contaminación exponencial con una menor precipitación anual y menor desplazamiento del UTC. Podríamos llamarla **"Ciudades con menor precipitación y mayor contaminación"**.

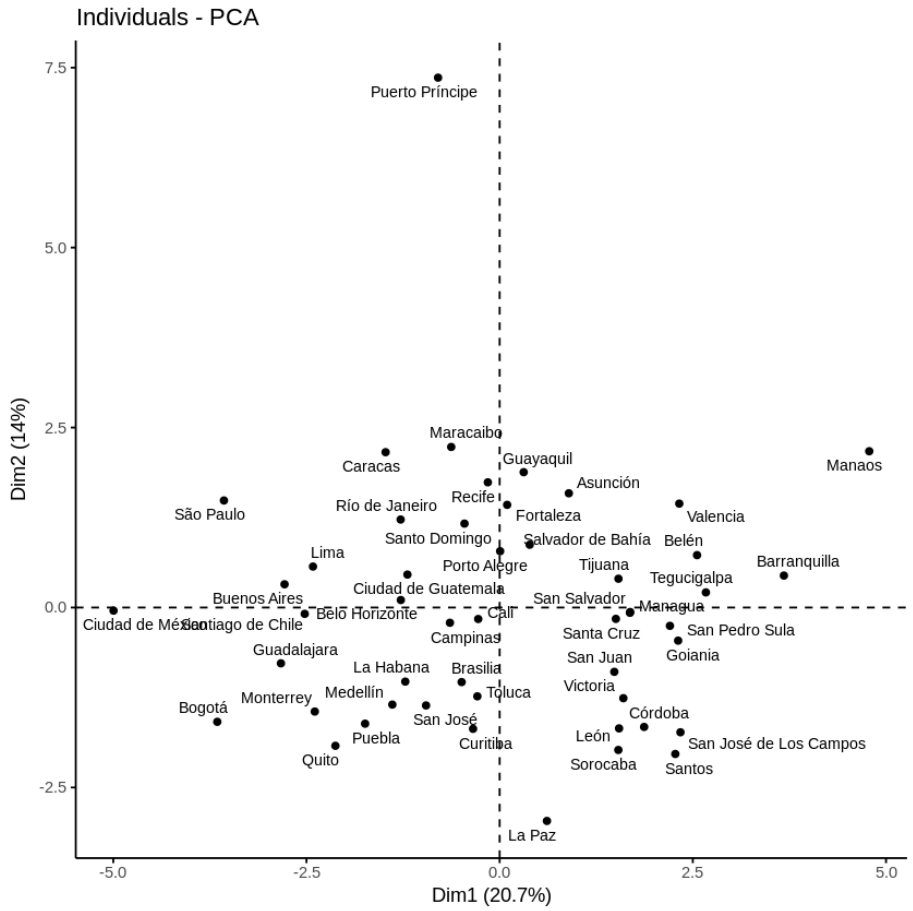
Componente 8:
- Cargas positivas altas: Tasa de homicidios, Tasa de criminalidad
- Cargas negativas altas: UTC Offset moderada

Este componente parece ubicar las ciudades de mayores tasas de homicidios y criminalidad anual en aquellas que se encuentran ubicadas en las locaciones con menor desplazamiento del UTC. Podríamos llamarla **"Ciudades con mayor criminalidad"**.

Componente 9:
- Cargas positivas altas: Precipitación media anual
- Cargas negativas altas: Salario promedio

Este componente parece asociar una mayor precipitación anual con menores salarios, sin embargo hay que recordar que primero que todo esta es una técnica descriptiva y la correlación no implica causalidad. Podríamos llamarla **"Ciudades con alta precipitación y salarios bajos".**

Analizando las 2 primeras componentes frente a las ciudades de la base de datos, podemos verla en el siguiente grafico:



Se visualizan las dos primeras componentes principales (Dim1 y Dim2), que explican el 20.7% y 14% de la varianza, respectivamente. Estos componentes agrupan las ciudades en función de diferentes indicadores socioeconómicos y urbanos, revelando patrones de desarrollo y condiciones ambientales.

La primera componente principal (PC1), que hemos denominado "Mal desarrollo Socioeconómico", distingue entre ciudades con alta tasa de desempleo, pobreza y curiosamente alfabetización, las ciudades ubicadas en el lado negativo del eje son aquellas que en contraste parecerían ser ciudades con un buen desarrollo económico, es decir las ciudades más grandes y desarrolladas, como Ciudad de México, Buenos Aires, Santiago de chile, São Paulo, entre otras, mientras que las ciudades en el eje contrario (positivo) serían aquellas con peor desarrollo económico pero una mayor alfabetización algunas de estas son Barranquilla, Tijuana, Manaos, entre otros.

La segunda componente principal (PC2) es más fácil analizarla mediante la rotación varimax, en este caso la llamamos "Ciudades costeras", muestra ciudades con climas cálidos y menor altitud, que están en la parte superior

del gráfico. Ejemplos de estas ciudades incluyen Barranquilla, Puerto Príncipe, Guayaquil y Maracaibo. En caso, es curioso que son las que parecieran tener menor número de universidades

Representación no estructurada:

Entendimiento del negocio:

Para la segunda parte del proyecto, nos enfocaremos en analizar las letras de las canciones más escuchadas en la última década, seleccionadas de la playlist "Top Latin Hits" de Spotify. Esta playlist, curada anualmente por la plataforma, refleja los éxitos más destacados de cada año y constituye una muestra representativa de las preferencias musicales de la audiencia latina.

La música no solo es entretenimiento; también es un reflejo de los cambios sociales, culturales y políticos de la región. Durante el período 2010-2019, géneros como el reggaetón, la cumbia y el pop han predominado en estas listas, impulsados por artistas latinos que han alcanzado una popularidad global. Para capturar esta diversidad y entender las temáticas recurrentes en la música popular, seleccionaremos las 10 primeras canciones de la playlist mencionada para cada año, sumando un total de 100 canciones.

En cuanto a la estructura de una canción, generalmente se compone de varias secciones que, en conjunto, crean una narrativa musical. Las partes principales son:

Verso: Expone la historia o mensaje de la canción. Suele variar en cada repetición para desarrollar la narrativa.
Coro: Es la parte más pegajosa y repetitiva. Resume el mensaje principal y es la sección que generalmente queda en la memoria del oyente.
Pre-coro: Sección que precede al coro y genera anticipación, conectando el verso con el coro.
Puente: Ofrece un contraste con el resto de la canción, rompiendo la monotonía y añadiendo dinamismo.
Outro: Cierra la canción, a veces repitiendo el coro o una variación del mismo.

Entender esta estructura nos permite descomponer las letras para un análisis más preciso, identificando no solo los temas dominantes, sino también cómo se distribuyen a lo largo de la canción.

Limpieza de datos: Inicialmente se cargaron las letras de las canciones desde un archivo Excel. En este paso se realizaron las siguientes acciones:

1. Eliminación de caracteres especiales y números: Se eliminan elementos no textuales que podrían interferir con el análisis de texto (como signos de puntuación o números).
2. Conversión a minúsculas: Para garantizar que el procesamiento sea uniforme, todas las letras se convirtieron a minúsculas.
3. Eliminación de stopwords: Utilizando una lista de palabras vacías en español (stopwords), eliminamos palabras comunes que no aportan valor semántico al análisis, como "el", "la", "que", etc. Esto permite que los modelos de tópicos identifiquen términos más relevantes.

Transformación de los datos: Para aplicar modelos de extracción de tópicos, fue necesario convertir las letras de canciones en representaciones numéricas. Utilizamos el modelo de TF-IDF (Term Frequency-Inverse Document Frequency), que asigna un peso a cada término en función de

su frecuencia relativa dentro de un documento y su frecuencia inversa en la colección de documentos. Este enfoque permite que las palabras con mayor relevancia en un documento específico, pero menos comunes en general, obtengan un mayor peso.
Aplicamos el método TfidfVectorizer de sklearn, con un umbral de frecuencia máxima del 85% (max_df=0.85) y mínima del 2% (min_df=2). El primer parámetro, indica que cualquier palabra que aparezca en más del 85% de las canciones será descartada, el motivo de esta exclusión es que tales términos son generalmente demasiado comunes y no aportan información diferenciadora para identificar temas. El segundo parámetro, establece que solo se incluirán términos que aparezcan en al menos dos canciones, si una palabra aparece únicamente en una canción, es probable que se trate de un término específico o de baja relevancia en el análisis global.

Embeddings

Para transformar las descripciones textuales de las canciones en vectores numéricos de alta dimensión, se utilizó el modelo 'multilingual-e5-large-instruct' de Sentence Transformers. Este modelo fue seleccionado debido a su destacado desempeño en el Massive Text Embedding Benchmark (MTEB) Leaderboard, lo que garantiza la calidad y representatividad de las embeddings generadas.

El proceso de generación de embeddings incluyó la combinación de las columnas mencionadas para formar una entrada textual única por canción, seguida de la transformación de estos textos en vectores numéricos mediante el modelo pre entrenado.

2. Estructura de la base de datos de canciones:

Artista: Nombre del artista o banda. Aunque podría almacenarse como un dato VARCHAR, puede haber variantes en la representación de los nombres.

Año: El año en que la canción fue lanzada. Es un campo numérico.

Canción: Título de la canción. Similar al campo "Artista", puede tener formatos diferentes, símbolos o caracteres especiales.

Playlist: Las canciones pueden estar asociadas con diferentes listas de reproducción. Esto implica una relación flexible y no necesariamente estructurada, ya que una canción puede aparecer en múltiples playlists.

Letra: Este campo es un texto largo que contiene la letra de la canción. Las letras no siguen una estructura clara y pueden variar en longitud, formato e incluso tener símbolos y caracteres especiales.

3. Modelado:

a. Representación no estructurada:

Seleccionamos técnicas de análisis de temas basadas en descomposición matricial y distribución de probabilidades para identificar los tópicos presentes en las letras de las canciones. Para el proyecto los modelos seleccionados fueron NMF (Non-negative Matrix Factorization) y LDA (Latent Dirichlet Allocation), con el objetivo de comparar sus resultados, pero ¿por qué elegimos estas dos?
Por un lado, es útil porque descompone la matriz de palabras (TF-IDF) en dos componentes: uno que representa las canciones y otro que representa los temas. Una de sus ventajas es que el

resultado es más fácil de interpretar, ya que siempre trabajamos con valores positivos, lo que ayuda a tener una idea clara de cuántas veces aparece un tema en las canciones. LDA, por otro lado, es un modelo más probabilístico, donde cada canción es modelada como una mezcla de temas, y cada tema es representado como una distribución de palabras.

Para construir los modelos, tomamos la matriz TF-IDF que habíamos generado previamente. Esta matriz resume qué tan importantes son ciertas palabras en las canciones, teniendo en cuenta que las palabras muy comunes o poco representativas (como “el”, “la”) se filtran. Luego, tanto para NMF como para LDA, seleccionamos un número de 10 temas (esto lo indicamos en los parámetros n_components=10 para NMF y n_topics=10 para LDA). Elegimos 10 porque pensamos que sería un buen equilibrio: nos permitiría encontrar una variedad de temas sin que los modelos se volvieran demasiado complicados o difíciles de interpretar. Los modelos ajustaron sus parámetros con los datos, lo que nos permitió extraer las palabras más importantes para cada tema. Luego, esas palabras nos dieron una idea de qué trata cada grupo de temas.

Embeddings:
Para transformar las descripciones textuales de las canciones en vectores numéricos de alta dimensión, se utilizó el modelo 'multilingual-e5-large-instruct' de Sentence Transformers. Este modelo fue seleccionado debido a su destacado desempeño en el Massive Text Embedding Benchmark (MTEB) Leaderboard, lo que garantiza la calidad y representatividad de las embeddings generadas.

El proceso de generación de embeddings incluyó la combinación de las columnas mencionadas para formar una entrada textual única por canción, seguda de la transformación de estos textos en vectores numéricos mediante el modelo preentrenado.

Reducción de Dimensionalidad con t-SNE
Dado que los embeddings generados son de alta dimensión, se aplicó la técnica t-SNE (t-Distributed Stochastic Neighbor Embedding) para reducir la dimensionalidad a dos dimensiones. Esta reducción facilita la visualización y el análisis de patrones en los datos, permitiendo una representación gráfica clara de las relaciones entre las canciones.

Visualización de los Resultados
Se realizaron dos tipos de visualizaciones para explorar la distribución de las canciones en el espacio reducido:

Visualización con Seaborn
Se generó un gráfico de dispersión utilizando Seaborn, donde cada punto representa una canción posicionada según sus componentes t-SNE, y se colorearon según el artista correspondiente. Esta visualización permite identificar agrupaciones naturales y observar cómo se distribuyen las canciones de diferentes artistas en el espacio.

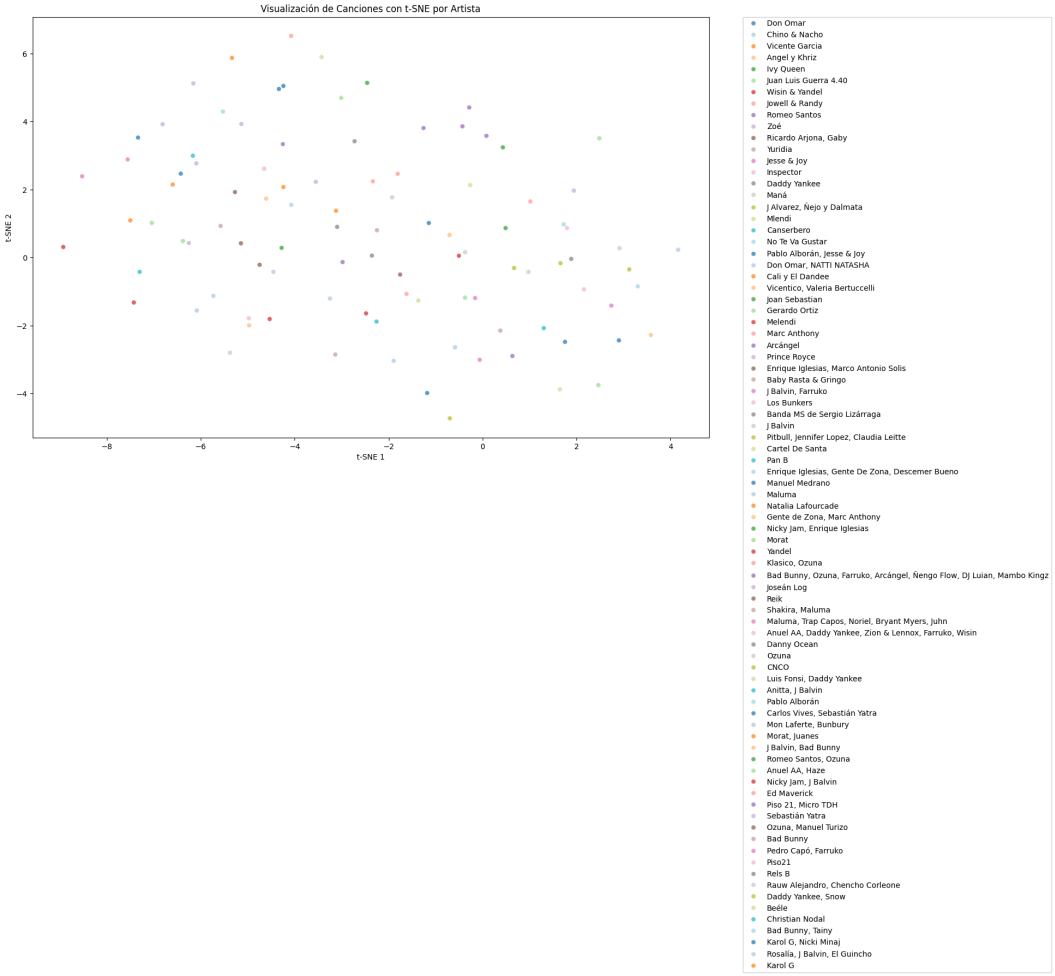


Figura 4: Visualización de Canciones con t-SNE por Artista (Seaborn)

Descripción: Gráfico de dispersión estático que muestra la distribución de las canciones en el espacio bidimensional de t-SNE, con colores diferenciados por artista.

Visualización Interactiva con Plotly

Adicionalmente, se creó una visualización interactiva utilizando Plotly, lo que permite una exploración dinámica de los datos. Esta herramienta facilita la identificación de relaciones y patrones específicos al permitir interactuar directamente con el gráfico, como acercar, alejar y obtener información detallada de cada punto al pasar el cursor.

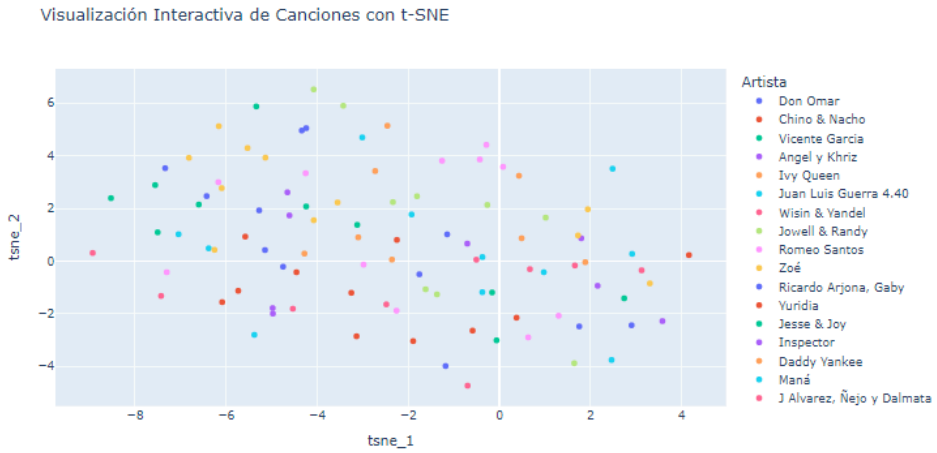


Figura 5: Visualización Interactiva de Canciones con t-SNE (Plotly)

Descripción: Gráfico interactivo que permite explorar la distribución de las canciones en el espacio de t-SNE, con capacidad de mostrar información adicional al interactuar con los puntos.

Analisis de la visualizacion:

Observaciones Generales

1. **Distribución de Canciones en el Espacio t-SNE**:
    - La distribución de los puntos es bastante dispersa, sin formar conglomerados muy densos. Esto sugiere que las canciones de diferentes artistas no están claramente agrupadas de manera consistente.
    - Algunos puntos se agrupan de forma más cercana, lo que podría indicar canciones que comparten similitudes textuales o temáticas, mientras que otros están más dispersos, indicando mayor variabilidad entre ellos.
2. **Presencia de Agrupaciones Locales**:
    - Se observan pequeñas agrupaciones de puntos que pueden indicar canciones del mismo artista que comparten características comunes (como temas similares en las letras). Estas agrupaciones tienden a aparecer de forma esporádica en todo el gráfico.
    - Sin embargo, la separación entre estas agrupaciones no es particularmente marcada, lo que puede indicar que las diferencias entre las canciones de distintos artistas no son tan significativas en el espacio semántico definido por los embeddings.

Análisis de Clustering

Además de la visualización, se realizó un análisis de clustering utilizando el algoritmo K-Means para agrupar las canciones en distintos clústeres. Se evaluaron diferentes números de clústeres utilizando el Método del Codo y el Silhouette Score para determinar la calidad y el número óptimo de agrupaciones.

Método del Codo

El Método del Codo implica graficar la inercia (suma de distancias al cuadrado de los puntos a sus respectivos centroides) en función del número de clústeres. El punto donde la reducción de la inercia comienza a disminuir de manera menos pronunciada sugiere el número óptimo de clústeres.

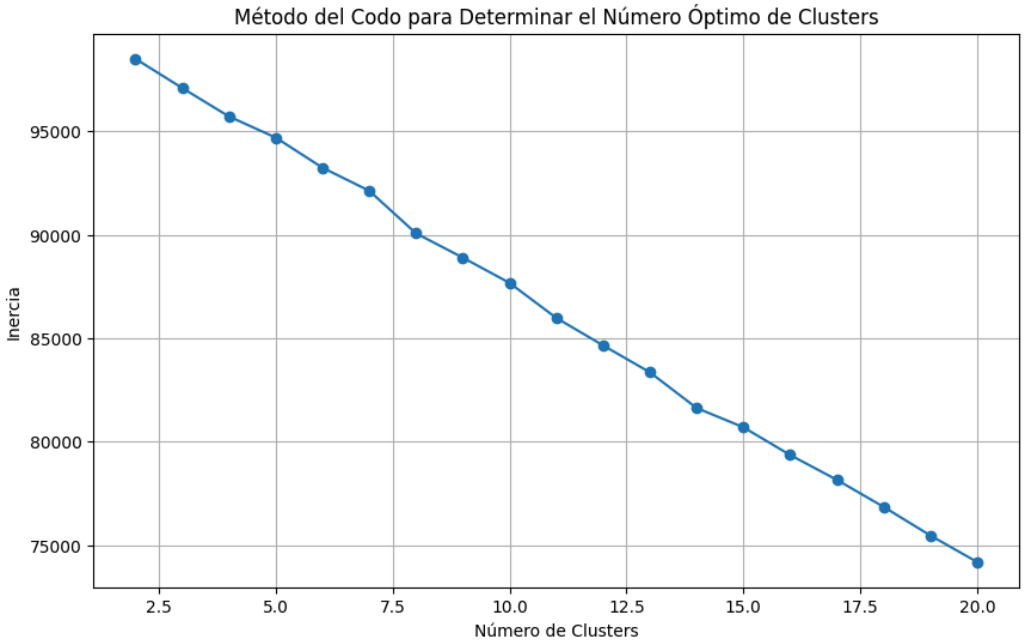


Figura 6: Método del Codo para Determinar el Número Óptimo de Clusters

Descripción: Gráfico de línea que muestra la inercia en función del número de clústeres, permitiendo identificar visualmente el "codo" que indica el número óptimo de agrupaciones.

Coeficiente de Silueta

El Silhouette Score mide qué tan bien está cada punto dentro de su clúster en comparación con otros clústeres. Los valores cercanos a +1 indican una buena cohesión y separación, mientras que valores cercanos a 0 o negativos sugieren clústeres superpuestos o mal asignados.

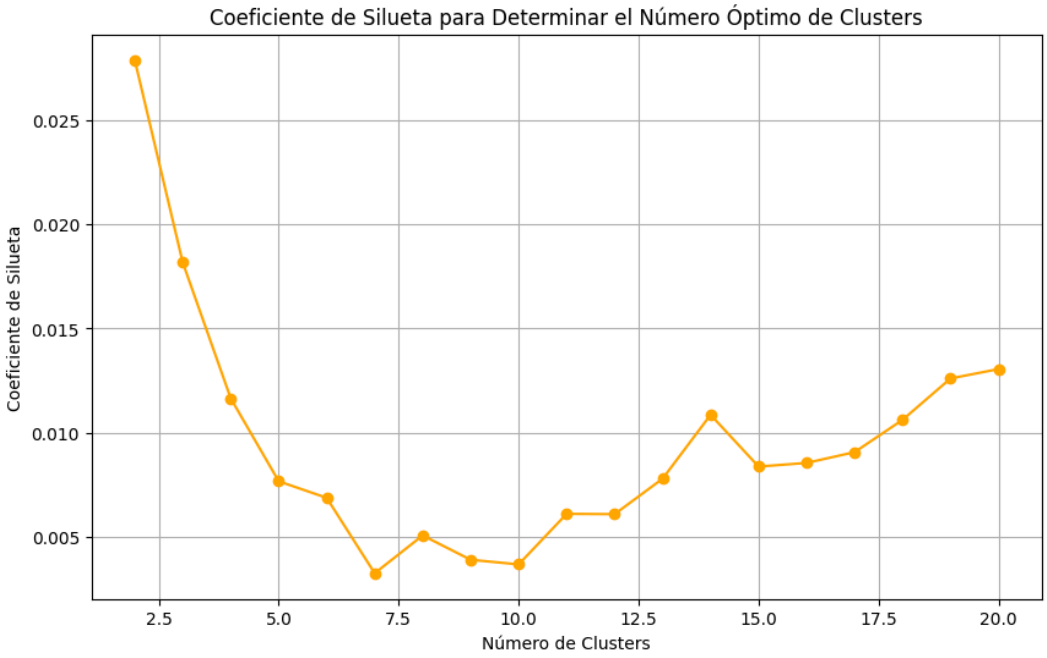


Figura 7: Coeficiente de Silueta para Determinar el Número Óptimo de Clusters

Descripción: Gráfico de línea que muestra el Silhouette Score en función del número de clústeres, facilitando la identificación del número de clústeres con mejor desempeño.

4. Evaluación

a. Representación no estructurada:

La evaluación de los modelos fue más interpretativa que técnica. Por un lado, con NMF vimos que ciertos temas se formaron alrededor de palabras más específicas y repetidas. Esto facilitó identificar categorías claras de las canciones, como relaciones amorosas, fiestas o reflexiones personales. Estos son los tópicos que encontramos y una posible interpretación:

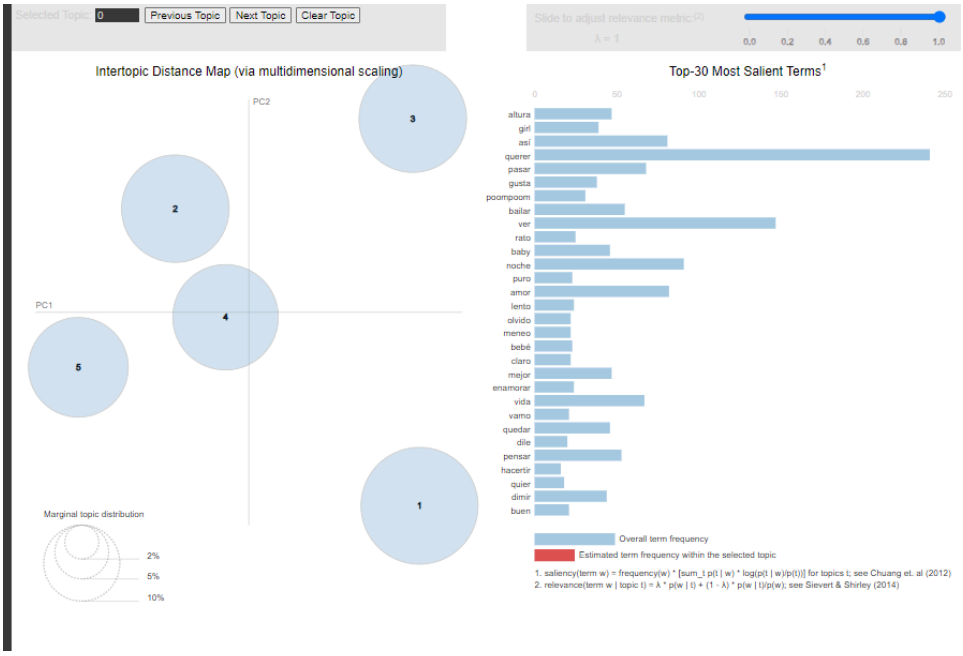
Evaluación de los tópicos de LDA:
Tópico 0: Contiene palabras como "amor", "prometo", "vida", "noche". Este tópico parece girar alrededor de promesas y esperanzas en relaciones amorosas.
Tópico 1: Aquí, palabras como "you", "contigo", "amor", "mirarte", mezclan inglés y español, probablemente reflejando la tendencia de muchas canciones latinas que incluyen palabras en inglés.
Tópico 2: Con términos como "besos", "enamore", "noches", se ve una vez más un enfoque en las relaciones románticas, pero esta vez con más énfasis en momentos físicos e íntimos como los besos y las noches compartidas.
Tópico 3: Palabras como "vamo", "miami", "puerto", pueden ser de un tema más festivo o vacaciones, aquí podrían estar reflejadas canciones de fiesta, típicas de géneros como reggaetón o música tropical.

Tópico 4: Contiene términos como "soñé", "corazón", "gente", lo que podría estar relacionado con canciones más reflexivas o nostálgicas.
Tópico 5: Las palabras "recuerdo", "pasando", "volvería", apuntan a un tema de rememoración y reflexión sobre el pasado, aquí podrían estar las canciones de “despecho”.
Tópico 6: Aquí vemos una mezcla de palabras "vida", "baby", "dime". Parece como una conversación íntima o coloquial entre amantes.
Tópico 7: Con términos como "quiero", "sé", "loco", se detecta un tono más directo y emocional, probablemente canciones sobre el deseo y el sentimiento de estar "loco" por alguien.
Tópico 8: Palabras como "ahhah", "nena", "musica".
Tópico 9: Contiene palabras como "foto", "flor", "invierno", lo que sugiere un tema más introspectivo, tal vez hablando de momentos o personas especiales con un enfoque poético.

Evaluación de los tópicos de NMF:
Tópico 0: Las palabras como "sé", "quieres", "baby", apuntan a un tema relacionado con el deseo o las relaciones, específicamente el diálogo entre amantes o personas que se quieren.
Tópico 1: Palabras como "contigo", "bailando", "feliz", parecen capturar un tema festivo, centrado en la alegría de bailar y compartir momentos felices, lo cual es típico en canciones de reggaetón o música latina de fiesta.
Tópico 2: Con palabras como "loca", "bailando", "novio", este tópico parece tocar temas relacionados con la fiesta y las emociones intensas que surgen en esos contextos, como enamorarse o sentirse "loco".
Tópico 3: Palabras como "quiero", "beso", "tenerte", apuntan a un tema romántico, pero con un enfoque más físico, probablemente explorando el deseo de estar con alguien.
Tópico 4: Aquí, con palabras en inglés como "you", "girl", "love", hay una mezcla cultural, posiblemente asociada con canciones que combinan inglés y español, lo que refleja la influencia de la música internacional en la música latina.
Tópico 5: Palabras como "amor", "cielo", "corazón", sugieren un tema muy romántico y poético, típico en baladas o canciones de amor clásicas.
Tópico 6: Con términos como "girl", "baby", "feliz", este tópico parece tener un enfoque más casual y juvenil, posiblemente hablando de enamoramientos o relaciones más ligeras y modernas.
Tópico 7: Palabras como "reír", "vida", "recuerdo", sugieren un tema nostálgico pero optimista, centrado en la vida, los recuerdos y los buenos momentos.
Tópico 8: Con palabras como "vida", "besos", "voz", se refleja una vez más un enfoque romántico, pero con un tono más poético y reflexivo.
Tópico 9: Palabras como "miami", "playa", "lento", sugieren un tema de vacaciones y fiesta, probablemente relacionado con canciones de verano o viajes.

Al observar los resultados de ambos modelos, es evidente que NMF produjo tópicos más claros y fáciles de interpretar. Cada grupo de palabras parece más centrado en un tema específico, lo que facilita sacar conclusiones sobre las temáticas principales de las canciones. Los temas predominantes incluyen el amor, las relaciones, la fiesta y las reflexiones sobre la vida.
Por otro lado, LDA generó tópicos más diversos, pero algunos de ellos presentan una mezcla más amplia de palabras que, aunque relacionadas, parecen menos cohesionadas.

Una mejor forma de ver los resultados es mediante el siguiente grafico:



Usar el siguiente link:https://colab.research.google.com/drive/13veZpesAY5c-4ZhYUkjnUnCFheKMbdPz?usp=sharing#scrollTo=PrtLYYRUDn_J

Evaluación del Proceso de Generación de Embeddings y Análisis de Clustering

En este apartado, se evalúa de forma integral cada etapa del proceso realizado, desde la generación de embeddings utilizando el modelo preentrenado hasta el análisis de los clústeres utilizando técnicas de reducción de dimensionalidad como t-SNE. Se examina el rendimiento del modelo, las limitaciones técnicas encontradas y las implicaciones de cada una de las técnicas utilizadas para comprender la estructura de los datos.

1. Generación de Embeddings

Modelo Utilizado:
Se empleó el modelo **'multilingual-e5-large-instruct'** de Sentence Transformers, seleccionado por su capacidad para crear representaciones vectoriales de alta calidad, especialmente en tareas multilingües. La elección de este modelo fue guiada por su buen desempeño en el **Massive Text Embedding Benchmark (MTEB) Leaderboard**, lo que asegura una buena generalización al trabajar con datos diversos como las letras de canciones.

Evaluación de la Calidad de los Embeddings:

- Los embeddings generados son representaciones numéricas de alta dimensión que capturan la semántica de las letras de las canciones, así como detalles relacionados con los artistas y títulos.
- La calidad de estos embeddings se refleja en la capacidad del modelo para agrupar canciones similares en espacios de menor dimensión. Sin embargo, la efectividad del

modelo también depende de la diversidad y el balance de los datos de entrada, como las letras de las canciones y los estilos de los artistas.

2. Reducción de Dimensionalidad con t-SNE

Uso de t-SNE para la Visualización:

La técnica **t-SNE (t-Distributed Stochastic Neighbor Embedding)** se utilizó para reducir la dimensionalidad de los embeddings y permitir una visualización en dos dimensiones. Esta técnica es efectiva para preservar las relaciones locales entre los puntos, mostrando cómo algunas canciones de ciertos artistas pueden agruparse debido a similitudes en su contenido textual.

Evaluación de la Visualización t-SNE:

- **Preservación de Estructura Local:** t-SNE es conocido por su capacidad para mantener la proximidad entre puntos cercanos, lo que permite identificar agrupaciones naturales a nivel local. Esto fue visible en algunos grupos de canciones que, al ser proyectadas en 2D, quedaron más cercanas entre sí.
- **Limitaciones de la Técnica:** A pesar de sus ventajas, t-SNE no preserva bien la estructura global de los datos, lo que significa que las distancias entre puntos lejanos en el gráfico no siempre reflejan su similitud real. Esto se evidenció en la dispersión de canciones de distintos artistas y la falta de clústeres definidos en la visualización.
- **Desempeño Dependiente de Parámetros:** El resultado de t-SNE depende de parámetros como la **perplexidad** y el número de iteraciones. En este análisis, se utilizaron valores estándar, lo que pudo influir en la forma final de la visualización. Explorar distintos parámetros podría mejorar la representación de los datos.

3. Análisis de Clustering

Uso de K-Means y Evaluación del Número de Clústeres:
Se implementó el algoritmo de **K-Means** para intentar identificar grupos de canciones similares. Se evaluaron distintos números de clústeres utilizando el **Método del Codo** y el **Silhouette Score** para determinar la mejor cantidad de agrupaciones.

Evaluación de la Calidad de los Clústeres:

- **Método del Codo:** Este método mostró una disminución suave de la inercia conforme se aumentaba el número de clústeres, sin un punto de inflexión claro. Esto indica que no hay una cantidad de clústeres que destaque por ser óptima, sugiriendo una falta de estructura clara en los datos.
- **Silhouette Score:** Los valores del Silhouette Score fueron consistentemente bajos, cercanos a cero, lo que implica que las agrupaciones formadas no son coherentes. Las canciones dentro de un mismo clúster no son significativamente más cercanas entre sí que a canciones de otros clústeres.
- **Interpretación de Resultados:** La ausencia de clústeres bien definidos puede ser un indicativo de que las características de las canciones (letras, títulos, artistas) no presentan patrones claros de separación en el espacio vectorial generado. Esto puede deberse a la diversidad de estilos y temas en las canciones, que diluye las diferencias entre los grupos.

4. Limitaciones y Consideraciones Técnicas

- **Pérdida de Información en la Reducción de Dimensionalidad:** Al pasar de un espacio de alta dimensión (embeddings) a uno de dos dimensiones con t-SNE, es inevitable que se pierda información relevante. Esto puede llevar a que relaciones importantes entre las canciones no se reflejen de manera precisa en el espacio reducido.
- **Impacto del Preprocesamiento:** El preprocesamiento de las letras y la combinación de las columnas para crear el texto de entrada al modelo puede influir en los resultados. La forma en que se normalizan los textos, se eliminan caracteres especiales y se estructuran las entradas es fundamental para la calidad de los embeddings generados.
- **Dificultad de Separar por Artistas:** La visualización y el análisis de clústeres sugieren que la separación de canciones por artista no es suficiente para generar grupos distintivos. Esto puede ser un reflejo de la variabilidad en la forma de escribir las canciones y los temas abordados por cada artista.