Tecnológico de Monterrey

Herramientas computacionales:
el arte de la analítica
(Gpo 570)

Actividad Evaluable:

Patrones con K-means

Jorge Alan Ramírez Elías A01701350

Cómo correr el programa

Para correr el programa se deberá clonar el repositorio con los archivos necesarios en el siguiente enlace https://github.com/JARE2001/Herramientas-computacionales-el-arte-de-la-anal-tica

Una vez instalado, deberá redireccionarse a la carpeta en consola y entrar a la carpeta act 3. Para correr el programa deberá ejecutar el comando 'python graficas.py'. Cabe mencionar que se requiere tener Python instalado en el computador de antemano.

Reporte

En este reporte se estará analizando datos provenientes de Spotify que describen características sobre 195 canciones de esta misma plataforma. El autor que extrajo estos datos analizó las características de 100 canciones que le gustó y 95 que no le gustó. Con esto en mente, podemos proceder a realizar un mejor análisis con los datos recopilados, ver si hay patrones en ellos y, por supuesto, interpretarlos.

Este conjunto de datos se saco de https://www.kaggle.com/bricevergnou/spotify-recommendation?select=data.csv. El cual a su vez se obtuvo usando la API de Spotify.

Para iniciar a analizar los datos tenemos que entender las variables que se tomaron en cuenta para cada canción listada. De esta manera, al momento de interpretar los datos y gráficas nos será más fácil saber por qué actúan de la manera que actúan.

En el conjunto de datos se tiene 195 registros, significando cada una de las canciones, con 14 variables para cada uno, que serán las diferentes características de las canciones. En la siguiente tabla se muestra mejor cada variable, el tipo de dato que se utiliza y una breve descripción de a qué se refiere la columna.

Variable	Tipo de dato	Descripción
danceability	float64	La bailabilidad describe qué tan adecuada es una pista para bailar en función de una combinación de elementos musicales que incluyen tempo, estabilidad del ritmo, fuerza del ritmo y regularidad general. Un valor de 0.0 es menos bailable y 1.0 es más bailable.
energy	float64	La energía es una medida de 0.0 a 1.0 y representa una medida perceptiva de intensidad y actividad. Por lo general, las pistas enérgicas se sienten rápidas, fuertes y ruidosas. Las características perceptivas que contribuyen a este atributo incluyen el rango dinámico, el volumen percibido, el timbre, la tasa de inicio y la entropía general.
key	int64	La clave en la que se encuentra la pista. Los números enteros se asignan a tonos utilizando la notación estándar de clase de tono. Por ejemplo, 0 = C, 1 = C♯/D♭, 2 = D, y así sucesivamente.
loudness	float64	El volumen general de una pista en decibeles (dB). Los valores típicos oscilan entre -60 y 0 db.
mode	int64	Indica la modalidad (mayor o menor) de una pista, el tipo de escala de la que se deriva su contenido melódico. Mayor está representado por 1 y menor es 0.
speechiness	float64	Detecta la presencia de palabras habladas en una pista. Los valores superiores a 0,66 describen pistas que probablemente estén formadas en su totalidad por palabras habladas. Los valores entre 0,33 y 0,66 describen pistas que pueden contener tanto música como voz, ya sea en secciones o en capas, incluidos casos

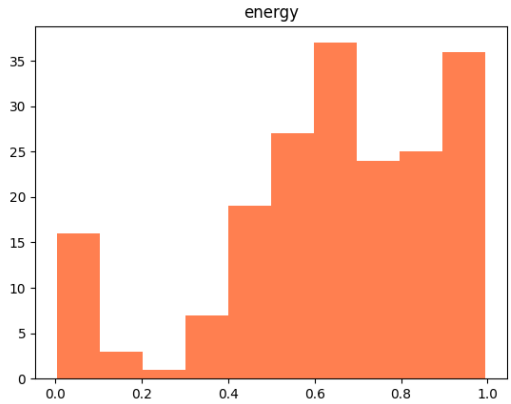
		como la música rap. Los valores por debajo de 0,33 probablemente representen música y otras pistas que no sean de voz.
acousticness	float64	Una medida de confianza de 0.0 a 1.0 de si la pista es acústica. 1.0 representa una alta confianza en que la pista es acústica.
instrumentalness	float64	Predice si una pista no contiene voces. Los sonidos “Ooh” y “aah” se tratan como instrumentales en este contexto. Cuanto más cerca esté el valor de instrumentalidad de 1.0, mayor será la probabilidad de que la pista no contenga contenido vocal. Los valores superiores a 0,5 pretenden representar pistas instrumentales, pero la confianza es mayor a medida que el valor se acerca a 1,0.
liveness	float64	Detecta la presencia de una audiencia en la grabación. Los valores de vivacidad más altos representan una mayor probabilidad de que la pista se interprete en vivo. Un valor superior a 0,8 proporciona una gran probabilidad de que la pista esté activa.
valence	float64	Una medida de 0.0 a 1.0 que describe la positividad musical transmitida por una pista. Las pistas con una valencia alta suenan más positivas, mientras que las pistas con una valencia baja suenan más negativas.
tempo	float64	El tempo general estimado de una pista en pulsaciones por minuto (BPM). En terminología musical, el tempo es la velocidad o ritmo de una pieza dada y se deriva directamente de la duración promedio del tiempo.
duration_ms	int64	La duración de la pista en milisegundos.
time_signature	int64	una firma de tiempo general estimada de una pista. El compás (medidor) es una convención de notación para especificar cuántos tiempos hay en cada compás.
liked	int64	1 para canciones que le gustan, 0 para canciones que no le gustan

Para las variables previamente elegidas, decidí analizar las columnas de energía y tempo de los datos recopilados.

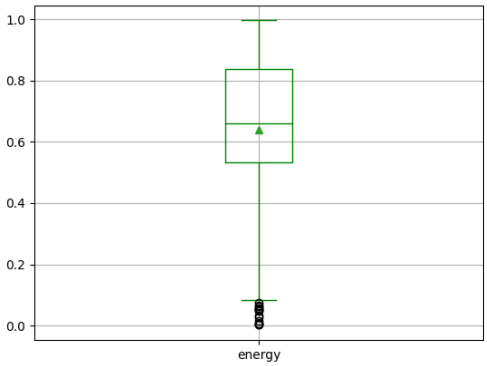
Energy

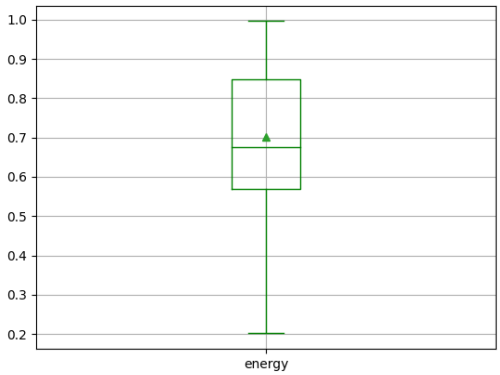
Como mencioné anteriormente en la tabla de variables, la energía es una medida de 0,0 a 1,0 y representa una medida perceptiva de intensidad y actividad. Entre más cercano sea el valor a 1.0, más energética será la canción.

Al observar los datos, se ve que la energía de las canciones varía entre 0.0024 y 0.996. En estos datos encontré que la ***desviación estándar*** es de **0.2600958155534138** por lo que los datos se encuentran medianamente dispersos cerca del 30% como bien se puede observar en las siguientes gráficas.



En esta gráfica podemos observar que la mayoría de los valores se encuentran entre 0.3 y 1.0 Por lo cual en las gráficas de abajo los valores debajo de 0.2 aparecen como outlier, fuera del rango de los cuartiles. Esta misma gráfica nos puede señalar que el autor que recopiló estos mismos datos de sus recomendaciones en Spotify tiene un mayor gusto por las canciones más energéticas y por esto mismo Spotify le recomienda canciones con más energía.

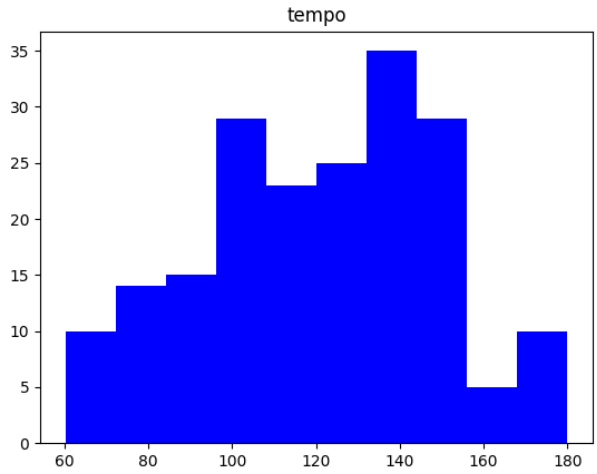




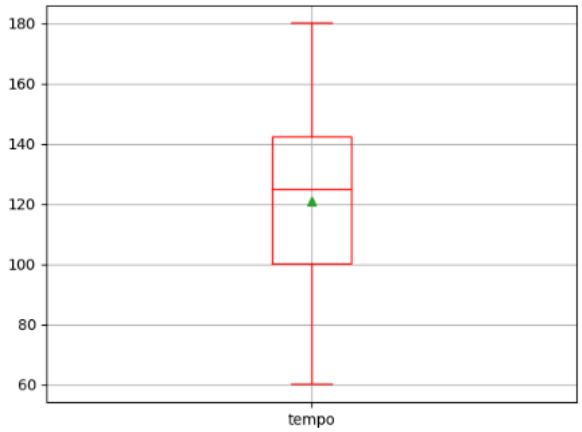
En estas gráficas podemos observar la **mediana** marcada por la línea del centro la cual se encuentra en un valor de **0.659** y la media o **promedio** que es de **0.6384314871794872**

Tempo

En los datos analizados el tempo representa la frecuencia de los pulsos por cada minuto (BPM). El tempo de las canciones en el data frame varió entre 180.036 y 60.17. En estos datos encontré que la ***desviación estándar*** es de ***28.08482882875693*** por lo que los datos se encuentran poco dispersos cerca del 10% de desviación del valor promedio.

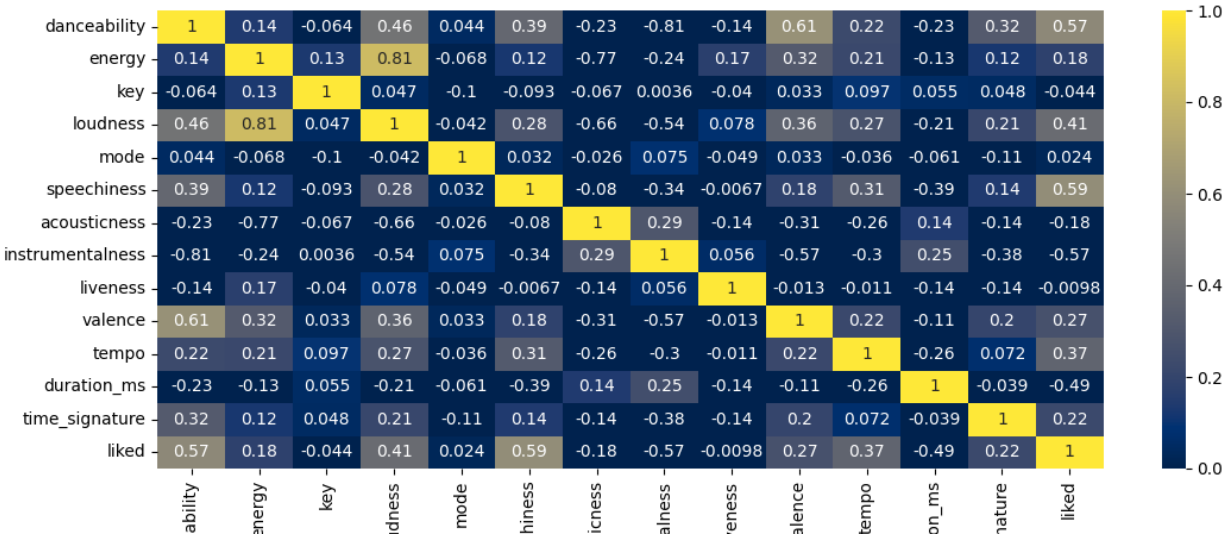


En la gráfica anterior se observa como mencionado anteriormente que los datos se encuentran en su mayoría en el centro por lo que la desviación estándar es pequeña. Esto igual nos dice que el tempo más repetido en las canciones es de 130 a 140, y podemos asumir que al oyente de estas canciones le suelen gustas más canciones con un ritmo más veloz a diferencia de las que tienen un tempo menor.



En esta grafica visualizamos la media y la mediana. La **mediana** representada por una línea se encuentra en **124.896.** La media/promedio es de **121.08617435897436.**

Correlaciones



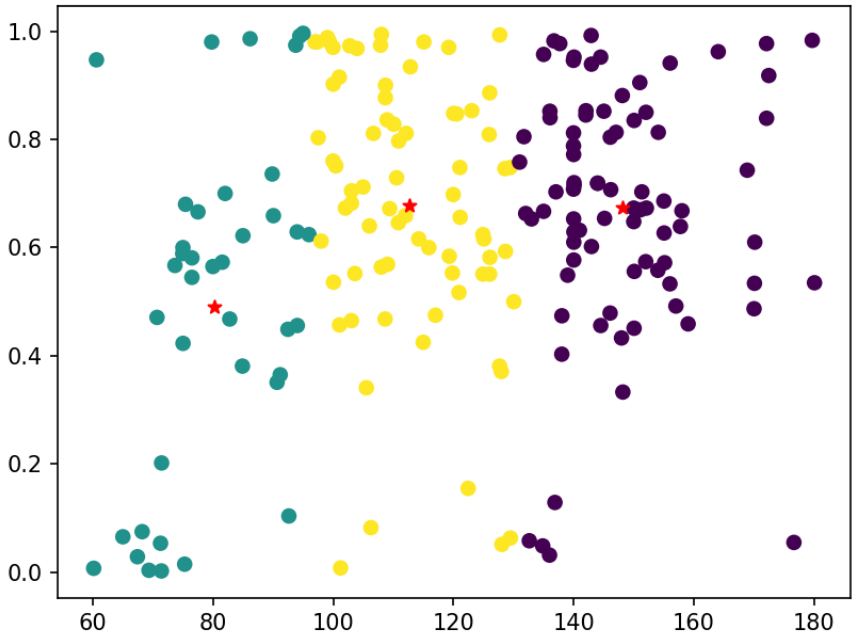
La gráfica anterior muestra la correlación entre las diferentes variables. Una de las más significativas se encuentra entre las variables de ‘energy’ y ‘loudness’, las cuales tienen una correlación de 0.81 y nos dicen que cuando una canción tiene un volumen alto, su energía también vendrá relacionada y será mayor.

Por otro lado, también podemos notar el caso de correlaciones negativas como lo es el de las variables de ‘acousticness’ y ‘energy’ las cuales tienen una correlación de -0.77 y nos dice que entre más acústica sea una canción, menos energía tendrá la misma. Esto hace sentido debido a que normalmente las canciones acústicas tienen a ser más calmadas y con tonos más relajantes, a lo contrario de como sería una canción energética.

Así como esta relación otra relación inversa(negativa) es la que hay entre ‘acousticness’ y ‘danceability’. La correlación tiene un peso de -.81 por lo que es una correlación bastante fuerte

Por ello es de mucho valor notar y analizar el mapa de calor, pues nos puede decir mucho sobre cómo las variables interactúan y se afectan entre ellas mismas.

KMEANS



En esta grafica podemos observar la relación entre las 2 variables que escogí la energía y el tempo. Como visto antes realmente la correlación de estas variables es baja de tan solo 0.21. Por esto mismo los puntos se encuentran muy dispersos por toda la gráfica. Sin embargo, yo decidí que el número de centros K fuera 3 ya que es la manera más lógica que encuentro de explicar las agrupaciones de datos. La gráfica se divide en 3 colores el primero azul representando en su mayoría a canciones con un a energía media baja y un tempo bajo. En segundo viene el grupo amarillo que representa canciones con una energía media alta y un tempo medio. Finalmente, el 3er grupo morado representa a aquellas canciones con las energías y los tempos más altos.

Si se usara menos centros muchos de los datos se encontrarían muy lejos de los centros y por lo tanto no habría una interpretación representativa. En el caso de generar más grupos se crearían categorías demasiado específicas y que crearían confusión en la relación entre los datos. Considero que la forma en la que lo grafique representa correctamente una relación baja pero positiva entre la energía y el tempo de la canción.

Los centros se encuentran divididos de forma casi exacta en 3 lo cual indica que los datos están correctamente segmentados. Si estos se encontraran muy juntos estaríamos dividiendo el mismo grupo en varios de más.

El que haya mucho outliers en el análisis genera que una porción considerable de los datos se encuentre fuera de los centros lo cual indica una correlación muy débil.