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

Reporte Datos Spotify

Paulina Cardoso Fuentes A01701490

Enero 2022

En este conjunto de datos se muestran datos de 195 canciones de Spotify en las que se analizan diferentes puntos de cada una de ellas. Este conjunto de datos se obtuvo de https://www.kaggle.com/bricevergnou/spotify-recommendation?select=data.csv en las que el autor Brice Vergnou agarró 100 canciones que le gustan y 95 que no le gustan de la playlist de Spotify Recommendations. Los datos analizados representan diferentes factores que tienen las diferentes canciones como bailabilidad, energía, acústica y más.

Al momento del análisis se obtuvieron 195 datos que representan las 195 canciones, y se analiza cada una de ellas en 14 diferentes variables. Las variables en este conjunto de datos son:

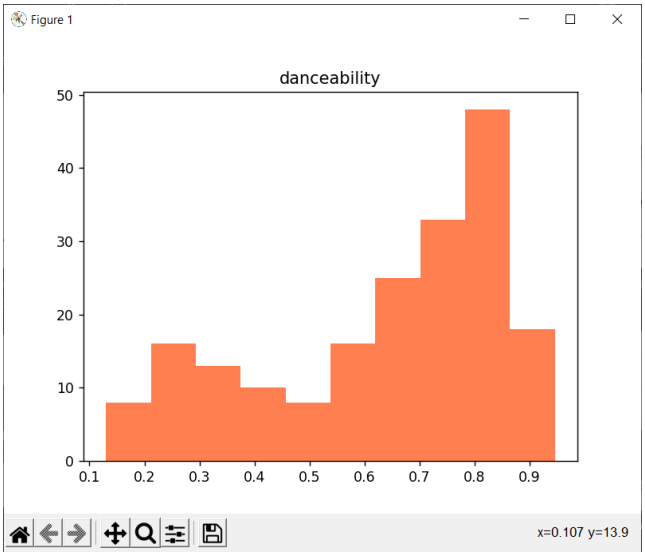
- **danceability:** número flotante que representa la bailabilidad de la canción
- **energy:** número flotante que representa la energía
- **key:** número entero que representa la nota de la canción usando una notación Pitch Class. Ejemplo: 0 = C, 1 = C♯/D♭, 2 = D…
- **loudness:** número flotante que representa lo fuerte del sonido de una canción en decibeles
- **mode:** número entero indicando la modalidad (mayor o menor) de una canción. Mayor representa 1 y menor es 0.
- **speechiness:** número flotante indicando la presencia de lo que se dice en una canción. Valores arriba de 0.66 indican que casi todo es hablado, entre 0.33 y 0.66 describe que la canción es ambas, entre música y cantada, y abajo de 0.33 indica que son canciones con más musica que letra.
- **acousticness:** número flotante que representa una medida entre 0.0 y 1.0 y dicta si una canción es acústica representando 1.0 como más acústico.
- **intrumentalness:** Número flotante que predice si la canción no contiene vocales, entre más cercano a 1.0, más probabilidad de que no tenga vocales.
- **liveness:** número flotante que detecta la presencia de audiencia en la grabación, entre mayor el número más probable que sea una canción presentada en vivo.
- **valence:** Número flotante que reoresenta una medida entre 0.0 y 1.0 describiendo una positividad musical de una canción.
- **tempo:** número flotante indicando el tiempo en beats por minuto de una canción
- **duration_ms:** número entero indicando la duración de la canción en milisegundos
- **time_signature:** número entero indicando el tiempo de firma de una canción
- **liked:** número entero indicando si la canción fue likeada o se haya gustado o no según el autor de los datos. 1 indicando que sí se gustó y 0 que no.

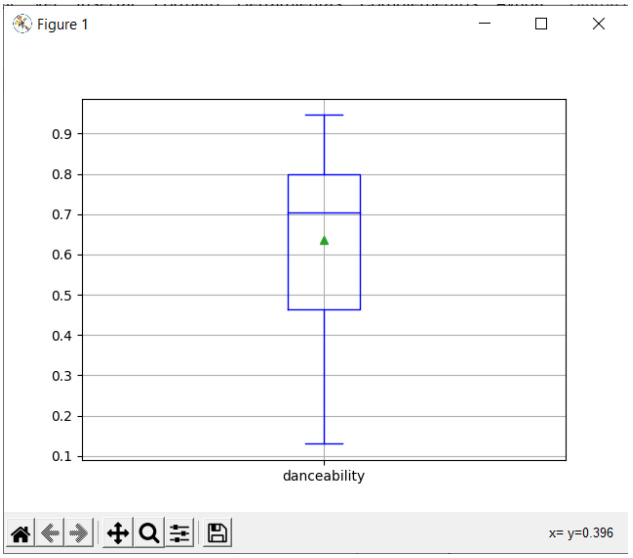
En este caso, elegí 2 variables que fueron ‘danceability’ y ‘key’ y analicé diferentes estadísticos en cada una de ellas.

En el primero de ellos, ‘danceability’ se obtuvieron resultados en un rango de 0.946 hasta 0.13. Hubo un promedio de 0.6366, mediana de 0.705 y desviación estándar de 0.2166. El número es una medida entre 0.0 y 1.0, 0.0 representando menos bailable y 1.0 más bailable. Por lo cual, con estos datos nos podemos dar cuenta que en general la mayor parte de las canciones que se tienen son un poco más de la mitad de bailables, tanto el promedio como la mediana tiene un valor arriba de 0.6 lo cual lo hace una canción bailable pero no tanto. Asimismo, por la desviación estándar nos podemos dar cuenta que los datos no son tan dispersos y nos dice que la mayor parte de las canciones son bailables.

Finalmente, para la segunda variable analizada, 'key', se obtuvo un rango entre 0 y 11, un promedio 5.4974, mediana de 6.0 y desviación estándar de 3.4152. Este valor lo que nos dice es en qué nota se encuentra la canción. Utilizando 0 = C = Do, 1 = C♯/D♭, 2 = D = Re, y así incrementando. Estos números nos indicaron que se tuvo un promedio de 5.49 que significa que el promedio de las notas de todas las canciones dieron que estuvieron en una nota 5, la cual es el equivalente a F, o bien, Fa. Mediana de 6, que sería la nota Fa sostenido mayor. Y la desviación estándar fue de 3.41, debido a que está más alejado de 0 nos podemos dar cuenta que los resultados en esta columna estuvieron más dispersos. Pude notar que en esta columna cada canción tiene diferentes resultados que me dice que no hay un patrón como tal.

Para tener una mejor visualización de los datos se realizaron diferentes gráficas que nos ayudan a poder tener una mejor idea de cómo se ven y cómo actúan entre ellos. Primero, para la columna 'danceability' que dicta qué tan bailable es una canción, se realizaron 2 gráficas, un histograma y un diagrama de cajas y bigotes, como se pueden observar en las siguientes figuras.

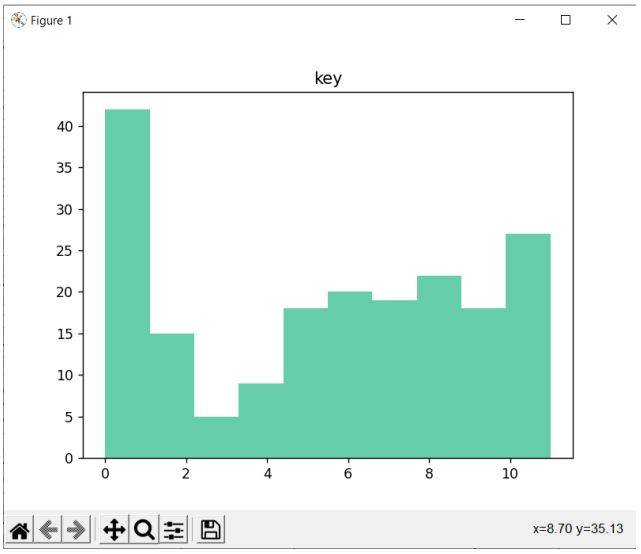


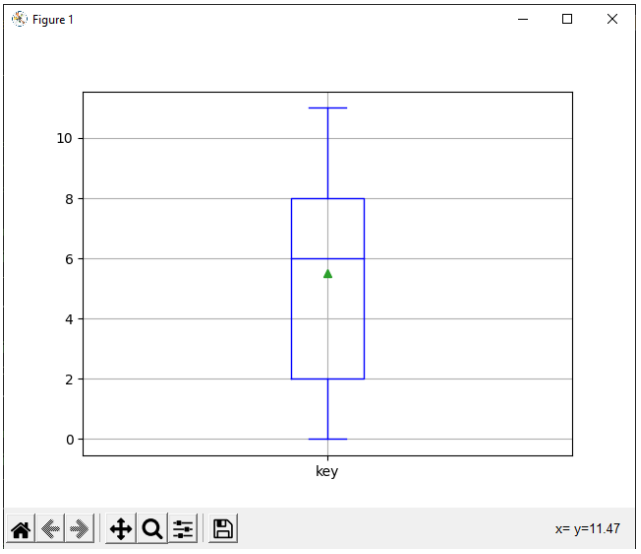


Con estas imágenes, podemos observar que, en la gráfica de la izquierda con el histograma la mayoría de los datos, en este caso canciones, tienen una calificación de 0.8 en bailabilidad, significando que son bailables. Si bien no todos los datos se encuentran en esta puntuación, sí podemos observar que la gran mayoría están entre 0.6 y 0.9, significando justamente que las canciones analizadas son más bailables en su mayoría

Por otro lado, con la gráfica de la derecha, el diagrama de cajas y bigotes observamos que el rango que va del cuartil 1 a la mediana es mucho más grande que el que va de la mediana al cuartil 3. Es decir, hay menos variación de datos entre 0.7 y 0.8 de bailabilidad en los datos que a diferencia de la variación de datos que hay entre casi 0.5 y 0.7. Y esto se puede ver y comprobar claramente al observar el histograma con lo que ya se mencionó anteriormente.

Asimismo, se realizaron estas mismas gráficas para la columna de 'key' para poder analizarla mejor. Estas gráficas fueron las siguientes:

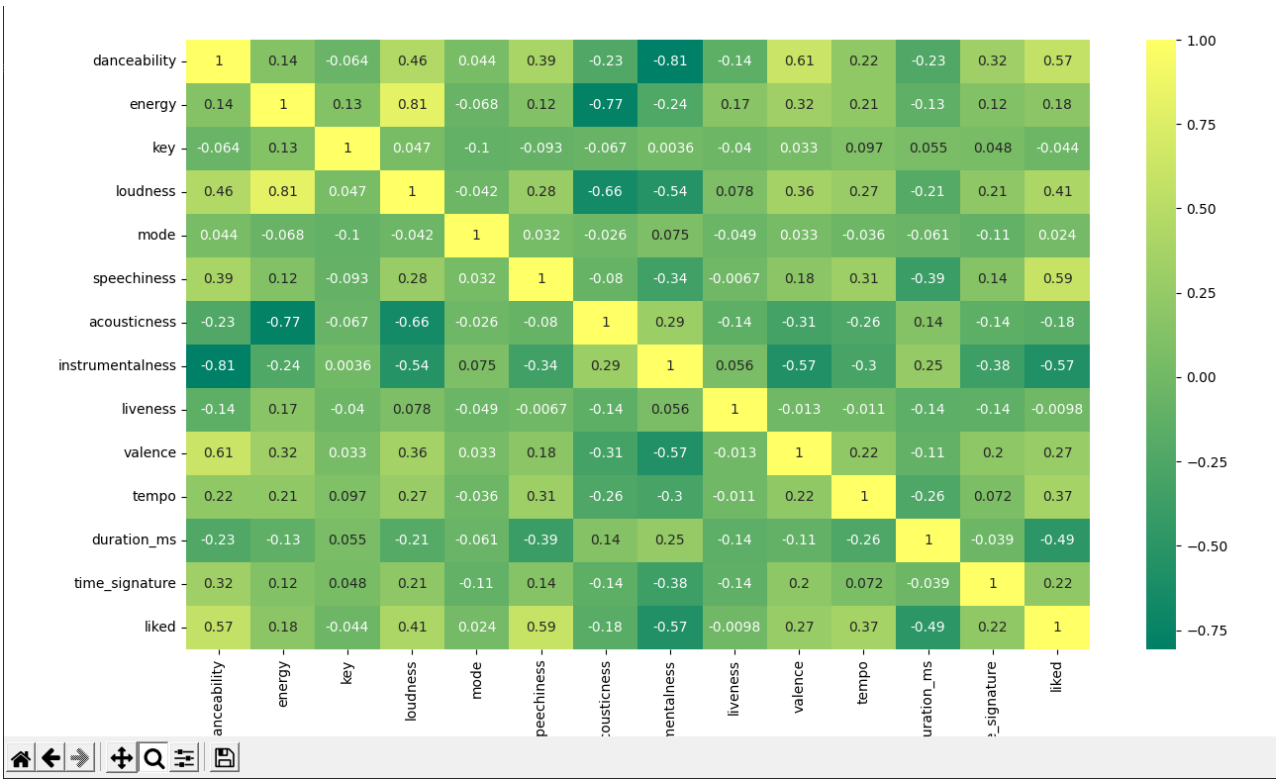




Nuevamente, al analizar la gráfica del histograma en esta ocasión para la columna 'key' podemos observar que la mayor cantidad de datos se encuentran en una nota 0, o bien Do. Sin embargo, a diferencia de 'danceability' la cantidad de datos con la cantidad de key es más variada. Significando que si bien la mayoría de datos está en una nota 0, se observa que los datos restantes están repartidos aproximadamente igual entre las notas de 5 a 11.

Con la otra gráfica observamos que, en nuestro diagrama de cajas y bigotes, se tiene una mayor diferencia entre las notas 2 a 6, que entre 6 y 8. Es decir que el rango entre el cuartil 1 a la mediana es mucho mayor que de la mediana al cuartil 3. Así como con las gráficas de 'danceability' esto también es notorio cuando observamos el histograma. Podemos ver que la cantidad de datos que se encuentran entre 6 y 8 tienen aproximadamente la misma altura, por lo que no difieren mucho. Y pasa lo opuesto para los datos que tienen una nota de 2 a 6.

Por último como gráficas analizadas está el mapa de calor, que nos muestra la correlación que tienen las variables entre ellas mismas. En el caso de este set de datos podemos observar lo siguiente:



Al observar algunas variables, nos damos cuenta que sí existen correlación entre algunas de ellas. Un ejemplo de esto está entre las variables de bailabilidad (‘danceability’) y instrumentación (‘instrmentalness’). Podemos notar que la correlación entre estas variables es de -0.81, esto debido a que entre menos instrumental sea la canción, será más bailable. Esto tiene sentido debido a que tendemos bailar aquellas canciones que tienen más vocales en ellos y menos instrumentos, no tendemos a bailar canciones con únicamente piano en ellas.

Otro ejemplo de las correlaciones entre estas variables lo podemos notar con las variables de energía y el volumen de la canción. También tiene sentido que entre mayor volumen tenga la canción, es decir, mayor número de decibeles, se tendrá más energía en la misma. Lo podemos observar fácilmente con las canciones de metal o rock.

Como estos 2 ejemplos hay muchos más en el que podemos notar la correlación entre las variables o de lo contrario cómo no se correlacionan. Un ejemplo de esto mismo es comparar la correlación entre las canciones que se gustaron (‘liked’) y la duración de la canción (‘duration_ms’), pues su correlación es de 0.22 con mucha razón porque lo que nos gusta de una canción es más su sonido y variables más generales sobre la melodía en lugar de su duración.

Con todo lo mencionado hasta ahora podemos decir que si bien hay variables que se correlacionan también existen muchas que no y al ver los datos completos y este mapa de calor podemos sacar más conclusiones de ellos como los que ya he mencionado.