TECNOLÓGICO DE MONTERREY, CAMPUS GUADALAJARA

BASES DE DATOS AVANZADAS

PROYECTO FINAL

EVALUACIÓN DEL ESTADO DE ÁNIMO CON BASE EN EL ANÁLISIS DE LA MÚSICA MÁS POPULAR DE LAS ÚLTIMAS DÉCADAS

INTEGRANTES:
Enrique Anaya Bovio - A01630317
Diego Alonso Martínez de Dios - A01228042

https://github.com/ebovio/MusicMiner

Prof. Rodolfo Rubén Álvarez González
Prof. Alberto de Obeso Orendain
03 de mayo de 2019

1. Introducción

El presente proyecto tiene como fin realizar un análisis del estado de ánimo en México a través de las décadas, partiendo desde 1957 a 2018. Para ello, evaluamos las 100 canciones más escuchadas cada año en nuestro país. Todos los códigos fuente del desarrollo del proyecto se encuentran dentro del repositorio:

https://github.com/ebovio/MusicMiner

1.1 Obtención de datos

Para el siguiente proyecto hemos utilizado la información de la siguiente página:

https://www.billboard.com/charts/hot-100

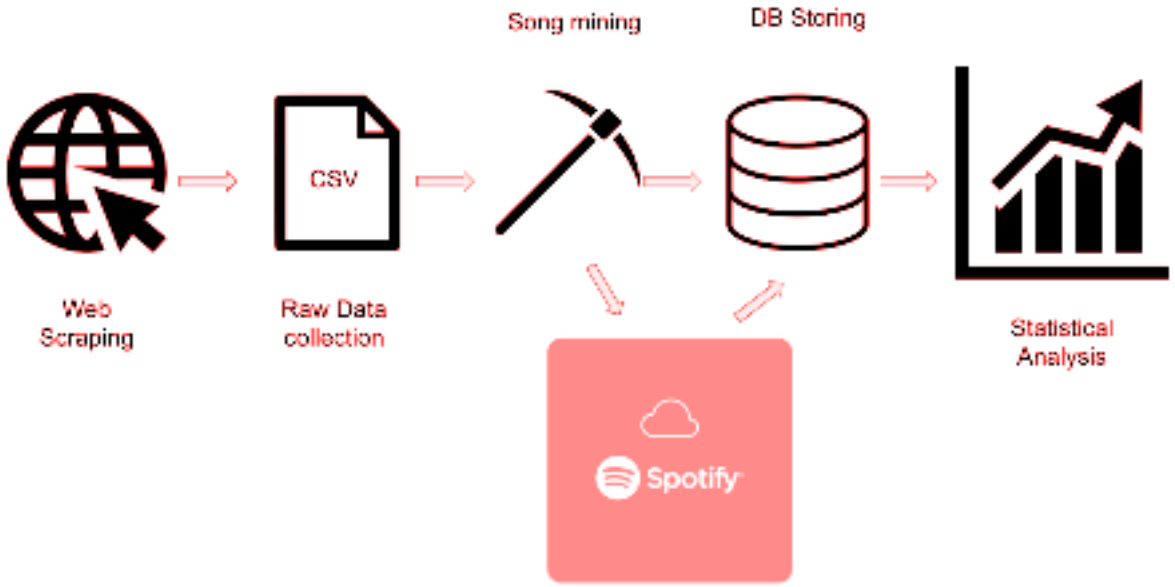
http://billboardtop100of.com/

Dichos datos fueron almacenados en un archivo .csv llamado 'Top_Songs_1940_2018.csv' que contiene el año, nombre y artista de cada canción.

1.2 Marco teórico

En definición de la página de Spotify para desarrolladores, la valencia se define como: 'Una medida entre 0.0 y 1.0 describiendo el positivismo de la canción. Canciones con mayor valencia suenan más positivas (feliz, contenta, eufórica) mientras canciones con menor valencia suenan más negativas (triste, deprimente, enojada)'.

1.3 Arquitectura



2. Proceso

Para el desarrollo del proyecto tomamos en consideración los valores de la valencia.

Para conocer e instalar las librerías y dependencias necesarias, referirse al punto 4.

2.1 Configuración

El programa utiliza una base de datos de MongoDB llamada ‘canciones’ y una colección llamada ‘col_name’. Además, al utilizar la API de Spotify es necesario obtener un token de acceso por parte del API en el sitio de Spotify for Developers.

2.2 Scrapping Song Titles

Para minar los datos simplemente ejecutamos el archivo ‘Miner.py’ dentro del mismo directorio que ‘Top_Songs_1940_2018.csv’.

El proceso tomará aproximadamente 30 minutos.

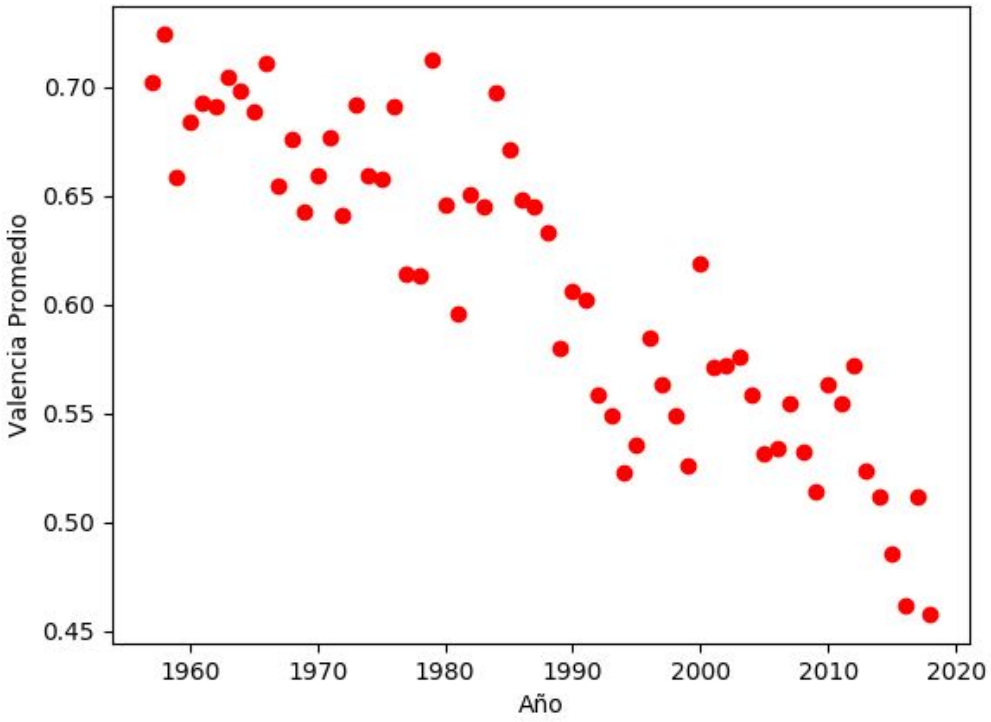
2.3 Procesamiento de datos

Una vez que se haya terminado el minado de datos, creamos las gráficas con la distribución de valencia por años al ejecutar el archivo ‘SongGrapher.py’.

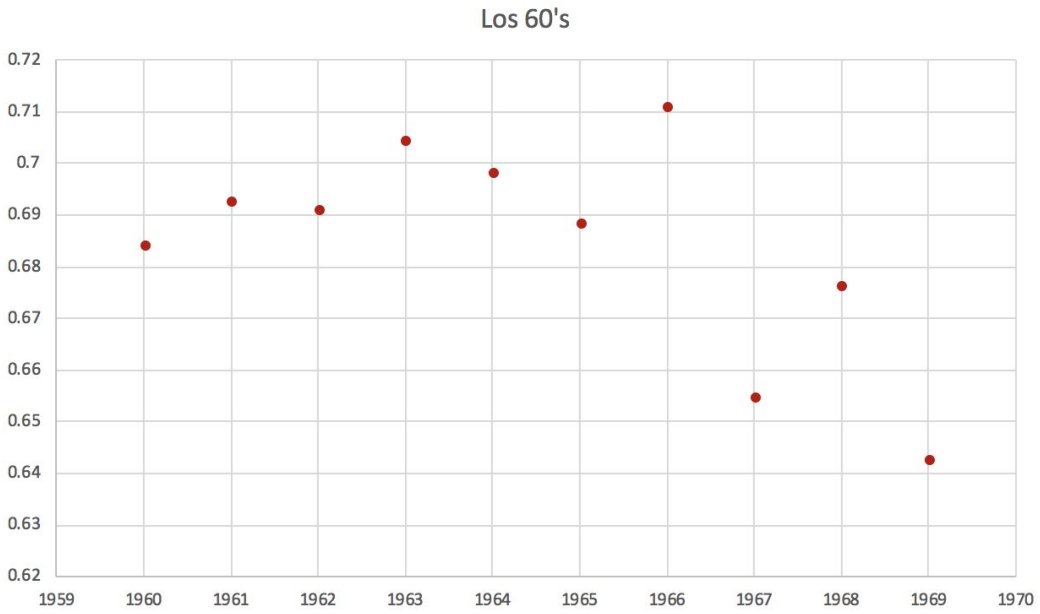
2.4 Limitantes del proyecto

No se cuenta con una base de datos para cotejar y comprobar que la información proporcionada por Spotify sobre el uso de la valencia como un indicador para describir el positivismo / negativismo de cada canción. Por lo tanto, confiamos en la información proporcionada por Spotify. Sin embargo, para corroborarla decidimos realizar encuestas cualitativas en las que evaluamos canciones que personas consideraban felices y tristes, a través de esta evaluación, pudimos observar que efectivamente los valores más altos correspondían a canciones consideradas felices y los más bajos a canciones tristes.

3. Resultados



3.1 Análisis de los datos por décadas



Media:

La valencia promedio de uno o cada año de la música escuchada durante la década de los 60 en México es de 0.68417916.

Mediana:

El 50% de la valencia de la música escuchada durante la década de los 60 en México es de más de 0.68956924 o menos.

Desviación estándar:

Lo dispersión promedio la valencia de la música escuchada durante la década de los 60 en México es de 0.02137209 con respecto a la media.

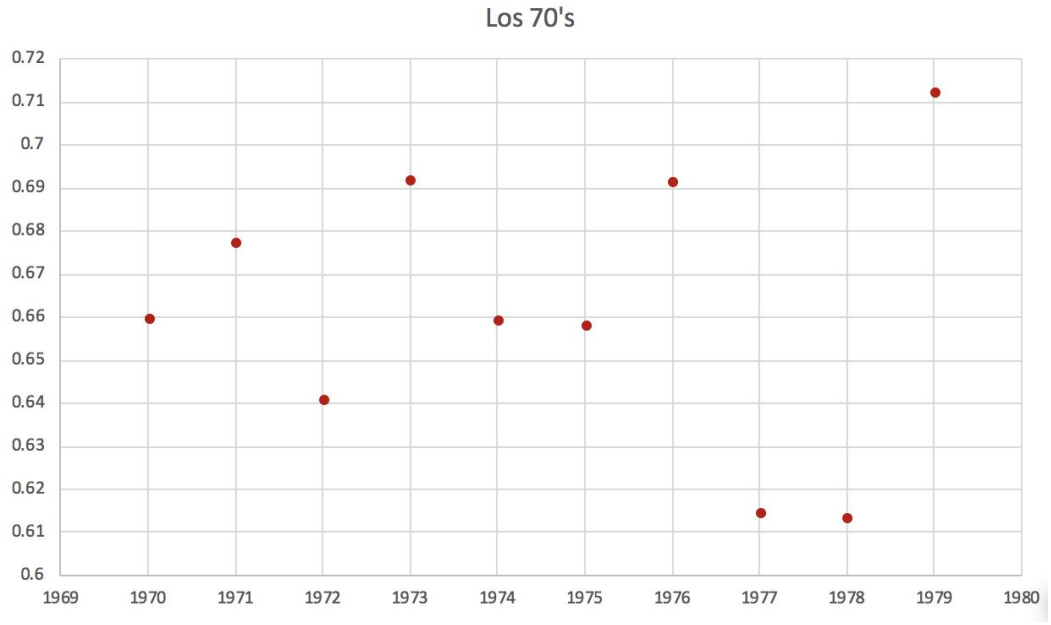
$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$ = 0.02137209

Coeficiente de variación:

$$Cv = \frac{\sigma}{\overline{x}} = \frac{0.02137209}{0.68417916}$$ = 0.031237563

Error estándar:

$$SE\overline{x} = \frac{S}{\sqrt{n}} = \frac{0.02137209}{\sqrt{10}}$$ = 0.0067585



Media:

La valencia promedio de uno o cada año de la música escuchada durante la década de los 70 en México es de 0.66161013.

Mediana:

El 50% de la valencia de la música escuchada durante la década de los 70 en México es de más de 0.65916337 o menos.

Desviación estándar:

Lo dispersión promedio la valencia de la música escuchada durante la década de los 70 en México es de 0.03268341 con respecto a la media.

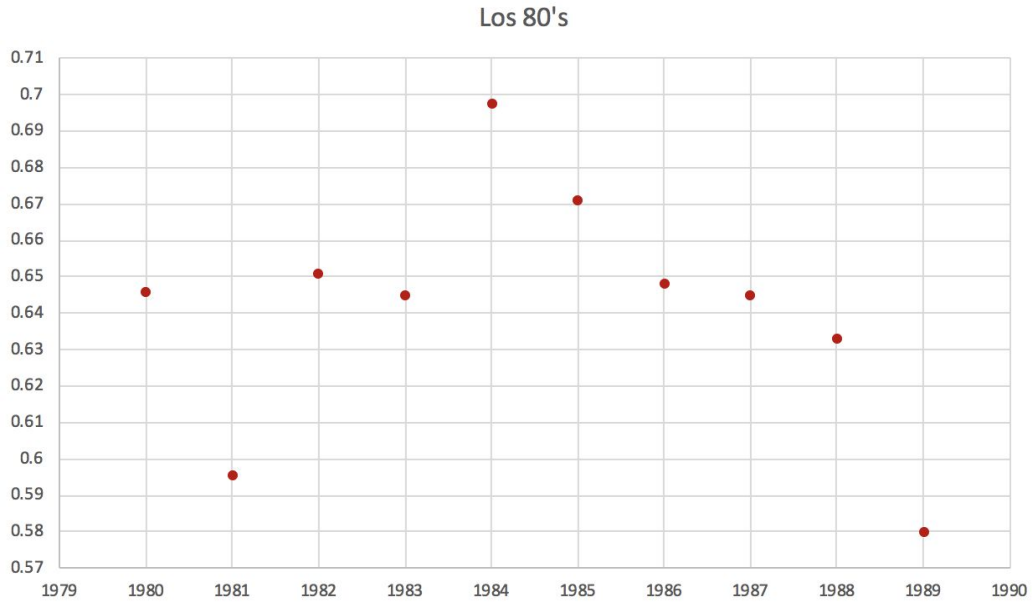
$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$ = 0.03268341

Coeficiente de variación:

$$Cv = \frac{\sigma}{\overline{x}} = \frac{0.03268341}{0.66161013}$$ = 0.049399802

Error estándar:

$$SE\overline{x} = \frac{S}{\sqrt{n}} = \frac{0.03268341}{\sqrt{10}}$$ = 0.010335619



Media:

La valencia promedio de uno o cada año de la música escuchada durante la década de los 80 en México es de 0.64114277.

Mediana:

El 50% de la valencia de la música escuchada durante la década de los 80 en México es de más de 0.64545561 o menos.

Desviación estándar:

Lo dispersión promedio la valencia de la música escuchada durante la década de los 80 en México es de 0.03359508 con respecto a la media.

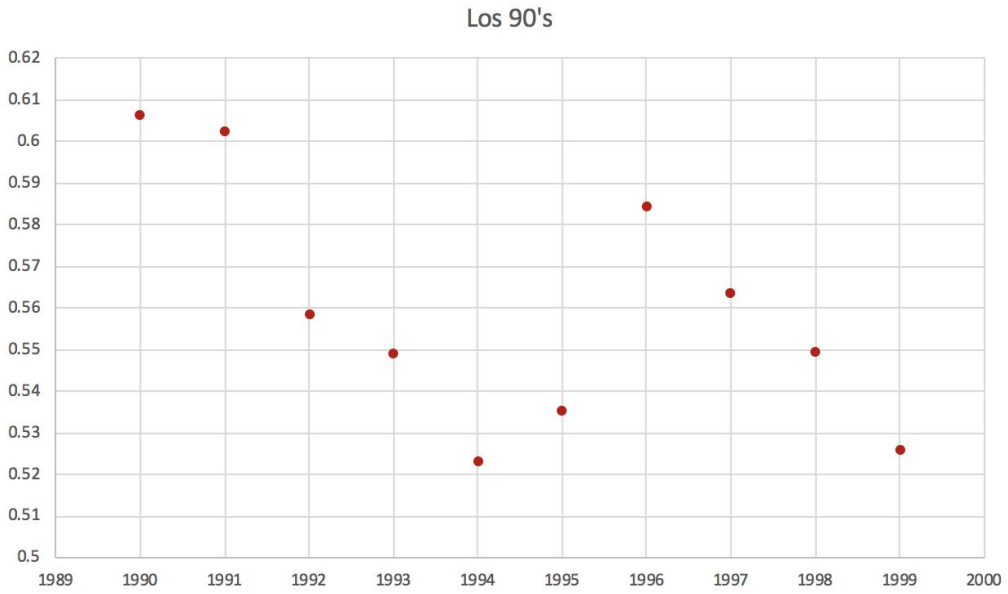
$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$ = 0.03359508

Coeficiente de variación:

$$Cv = \frac{\sigma}{\overline{x}} = \frac{0.03359508}{0.64114277} = 0.05239875$$

Error estándar:

$$SE\overline{x} = \frac{S}{\sqrt{n}} = \frac{0.03359508}{\sqrt{10}} = 0.010623697$$



Media:

La valencia promedio de uno o cada año de la música escuchada durante la década de los 90 en México es de 0.55971962.

Mediana:

El 50% de la valencia de la música escuchada durante la década de los 90 en México es de más de 0.553848 o menos.

Desviación estándar:

Lo dispersión promedio la valencia de la música escuchada durante la década de los 90 en México es de 0.02962908 con respecto a la media.

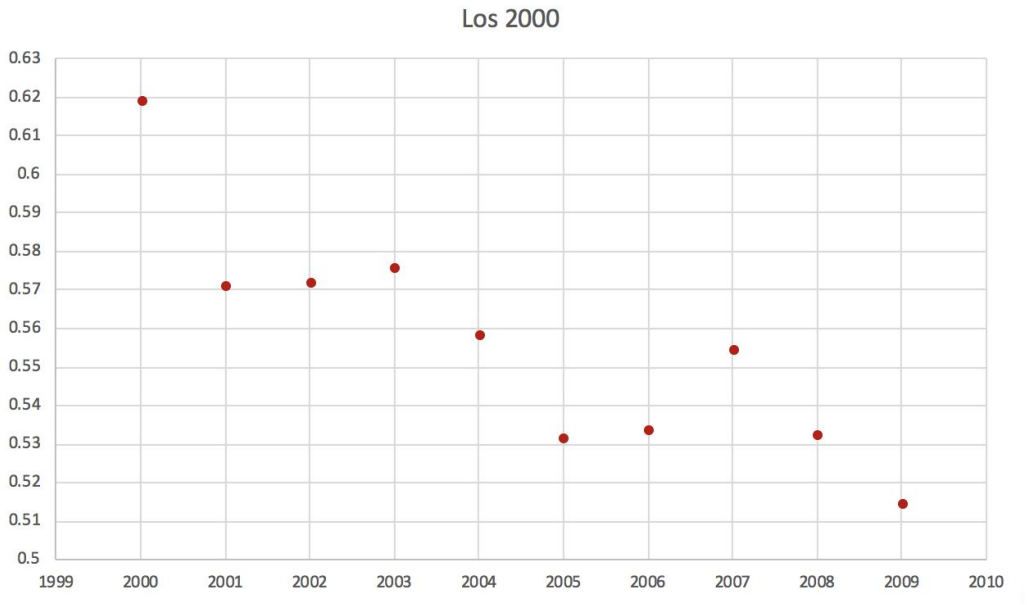
$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$ = 0.02962908

Coeficiente de variación:

$$Cv = \frac{\sigma}{\overline{x}} = \frac{0.02962908}{0.55971962}$$ = 0.052935575

Error estándar:

$$SE\overline{x} = \frac{S}{\sqrt{n}}$$ = $$\frac{0.02962908}{\sqrt{10}}$$ = 0.009369537



Media:

La valencia promedio de uno o cada año de la música escuchada durante los 2000 en México es de 0.55612071.

Mediana:

El 50% de la valencia de la música escuchada durante los 2000 en México es de más de 0.55635002 o menos.

Desviación estándar:

Lo dispersión promedio la valencia de la música escuchada durante los 2000 en México es de 0.02962908 con respecto a la media.

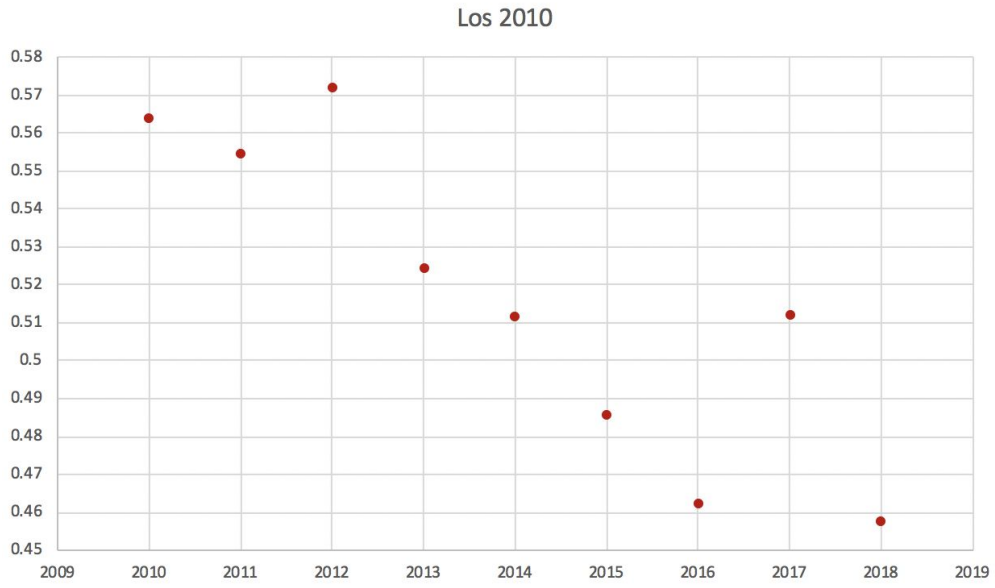
$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$ = 0.03027405

Coeficiente de variación:

$$Cv = \frac{\sigma}{\overline{x}} = \frac{0.03027405}{0.55612071}$$ = 0.054437911

Error estándar:

$$SE\overline{x} = \frac{S}{\sqrt{n}} = \frac{0.03027405}{\sqrt{10}}$$ = 0.009573495



Media:

La valencia promedio de uno o cada año de la música escuchada durante los 2010 en México es de 0.51582295.

Mediana:

El 50% de la valencia de la música escuchada durante los 2010 en México es de más de 0.51200159 o menos.

Desviación estándar:

Lo dispersión promedio la valencia de la música escuchada durante los 2010 en México es de 0.04212881 con respecto a la media.

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$ = 0.04212881

Coeficiente de variación:

$$Cv = \frac{\sigma}{\overline{x}} = \frac{0.04212881}{0.51582295} = 0.081673004$$

Error estándar:

$$SE\overline{x} = \frac{S}{\sqrt{n}} = \frac{0.04212881}{\sqrt{9}} = 0.014042936$$

3.2 Análisis general de los datos obtenidos

Tomando todos los datos arrojados por el programa realizamos un análisis general de la información obtenida desde 1957 hasta 2018. Obteniendo los siguientes resultados.

Media:

La valencia promedio de uno o cada año de la música escuchada durante las últimas 6 décadas en México es de 0.6089488.

Mediana:

El 50% de la valencia de la música escuchada durante las últimas 6 décadas en México es de más de 0.61372268 o menos.

Desviación estándar:

Lo dispersión promedio la valencia de la música escuchada durante las últimas 6 décadas en México es de 0.07053804 con respecto a la media.

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$ = 0.07053804

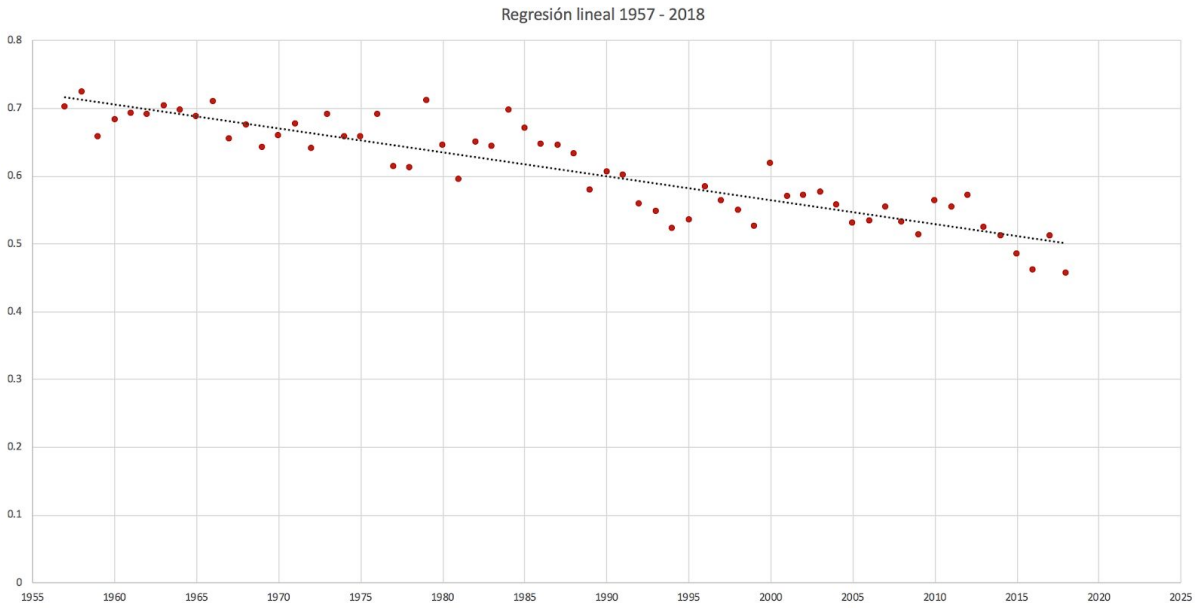
Coeficiente de variación:

$$Cv = \frac{\sigma}{\overline{x}} = \frac{0.07053804}{0.6089488} = 0.115835748$$

Error estándar:

$$SE\overline{x} = \frac{S}{\sqrt{n}} = \frac{0.07053804}{\sqrt{61}} = 0.00903147$$

Con base en los datos obtenidos de nuestro proyecto realizamos también una regresión lineal para hacer predicciones de los próximos años, tomando en cuenta la ecuación de la pendiente obtenida y = -0.0035x + 7.6143.



Si continuara dicho comportamiento para 2022 la el valor de la valencia disminuirá en un 4.35% y para 2030 dicho valor disminuiría un 7.62%.

4. Dependencias

Spotipy

https://spotipy.readthedocs.io/en/latest/

Pymongo

https://api.mongodb.com/python/current/

Numpy

https://www.numpy.org/

PyPlot

https://matplotlib.org/tutorials/introductory/pyplot.html