Autor: Rafael Hinojosa López

Matrícula: A01705777

Profesor: Gilberto Huesca Juárez

Materia: Herramientas computacionales: el arte de la analítica

Fecha: 12 - Enero - 2022

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Carga de Datos

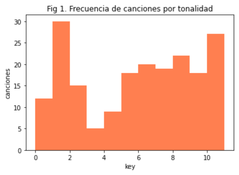
In [2]:
```
# 1. Cargar los datos
songs = pd.read_csv("../data/songs.csv")
```

In [3]:
```
# 2. Seleccionar dos columnas (key, tempo)
keys = songs['key']
tempos = songs['tempo']
```

Histogramas

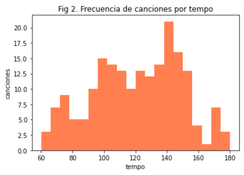
In [34]:
```
# 3a Histograma de key
songs.hist(column='key', bins=11, grid=False, orientation="vertical", color="coral")

plt.title('Fig 1. Frecuencia de canciones por tonalidad')
plt.xlabel('key')
plt.ylabel('canciones')
plt.show()
```



En la figura 1 se observa la distribución de frecuencias de canciones presentes en el dataset según la tonalidad de cada una. Se observa que aproximadamente 30 canciones están en el tono de C (key = 1), siendo este el tono más escuchado, y que 5 canciones están en el tono de D#/Eb (key = 3), siendo este el tono menos escuchado.

In [35]:
```
# 3b Histograma de key
songs.hist(column='tempo', bins=20, grid=False, orientation="vertical", color="coral")
plt.title('Fig 2. Frecuencia de canciones por tempo')
plt.xlabel('tempo')
plt.ylabel('canciones')
plt.show()
```

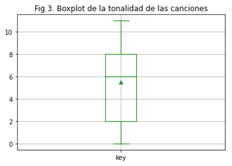


En la figura 2 se observa la distribución de frecuencias del de las canciones presentes en el dataset según *tempo* de cada una. Se observa que aproximadamente 20 canciones tienen un *tempo* entre 140 y 145 bpm, siendo este *tempo* el más escuchado. Por otra parte, el *tempo* menos escuchado corresponde al ubicado en el rango de 161 a 170 bpm, con aproximadamente una canción. A pesar de ser la menor frecuencia y tener un rango de *tempo* relativamente elevado con las demás canciones, el número de canciones escuchadas es mayor con rangos de *tempo* mayores.

Cajas y bigotes

In [36]:
```
# 4a. Gráficas de cajas y bigotes - Key
songs.boxplot(column="key", color = "green", showmeans=True)

plt.title('Fig 3. Boxplot de la tonalidad de las canciones')
plt.show()
```



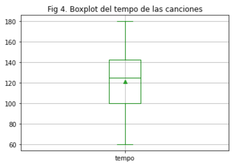
En la figura 3 se observen varios elementos: la media, la mediana, la desviación estándar, el mínimo y el máximo de las *keys* de las canciones.

La media está representada por un triángulo y la mediana por la línea horizontal dentro de la caja. La desviación estándar es la distancia que existe de la media a la parte alta de la caja, misma que es de la media a la parte baja de la caja.

Por último, el máximo valor es la línea horizontal fuera de la caja cuyo valor en *y* es 11 y el mínimo, la línea fuera de la caja cuyo valor en *y* es 0.

In [37]:
```
# 4b. Gráfica de cajas y bigotes - Tempo
songs.boxplot(column="tempo", color = "green", showmeans=True)

plt.title('Fig 4. Boxplot del tempo de las canciones')
plt.show()
```



Así como en la figura 3, en esta figura 4 se presentan los mismos datos pero de la variable *tempo*.

Podemos observar una caja más pequeña en altitud gracias al rango de valores que existen en el data set. De igual manera, la media y la mediana son muy cercanas.

Esto significa que probablemente que el número de canciones que tienen un *tempo* menor a la media es muy similar al número de canciones que hay con *tempo* mayor a la media.

Correlación y Mapas de Calor

In [25]:
```
# 5. Correlación de las variables numéricas
songs.corr()
```

Out[25]:

	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature	liked
danceability	1.000000	0.137188	-0.063906	0.455078	0.043759	0.388596	-0.234176	-0.807053	-0.137069	0.612344	0.223522	-0.232621	0.317096	0.569425
energy	0.137188	1.000000	0.130251	0.813567	-0.068308	0.122825	-0.772583	-0.241444	0.166508	0.319409	0.214905	-0.134527	0.123942	0.176179
key	-0.063906	0.130251	1.000000	0.046865	-0.103371	-0.093395	-0.066844	0.003597	-0.039622	0.033336	0.097240	0.054522	0.048344	-0.044406
loudness	0.455078	0.813567	0.046865	1.000000	-0.041678	0.279710	-0.664989	-0.538266	0.078093	0.363532	0.274462	-0.206334	0.207806	0.410774
mode	0.043759	-0.068308	-0.103371	-0.041678	1.000000	0.031953	-0.025709	0.075442	-0.048661	0.033409	-0.036270	-0.060965	-0.110739	0.023747
speechiness	0.388596	0.122825	-0.093395	0.279710	0.031953	1.000000	-0.079710	-0.343242	-0.006665	0.180708	0.313918	-0.388397	0.140325	0.591505
acousticness	-0.234176	-0.772583	-0.066844	-0.664989	-0.025709	-0.079710	1.000000	0.294320	-0.140988	-0.313806	-0.255097	0.138793	-0.142177	-0.179375
instrumentalness	-0.807053	-0.241444	0.003597	-0.538266	0.075442	-0.343242	0.294320	1.000000	0.055730	-0.572224	-0.299493	0.249683	-0.375199	-0.569440
liveness	-0.137069	0.166508	-0.039622	0.078093	-0.048661	-0.006665	-0.140988	0.055730	1.000000	-0.013004	-0.010555	-0.143966	-0.135409	-0.009797
valence	0.612344	0.319409	0.033336	0.363532	0.033409	0.180708	-0.313806	-0.572224	-0.013004	1.000000	0.218017	-0.114842	0.201111	0.268653
tempo	0.223522	0.214905	0.097240	0.274462	-0.036270	0.313918	-0.255097	-0.299493	-0.010555	0.218017	1.000000	-0.256250	0.071754	0.371202
duration_ms	-0.232621	-0.134527	0.054522	-0.206334	-0.060965	-0.388397	0.138793	0.249683	-0.143966	-0.114842	-0.256250	1.000000	-0.039078	-0.490651
time_signature	0.317096	0.123942	0.048344	0.207806	-0.110739	0.140325	-0.142177	-0.375199	-0.135409	0.201111	0.071754	-0.039078	1.000000	0.221479
liked	0.569425	0.176179	-0.044406	0.410774	0.023747	0.591505	-0.179375	-0.569440	-0.009797	0.268653	0.371202	-0.490651	0.221479	1.000000

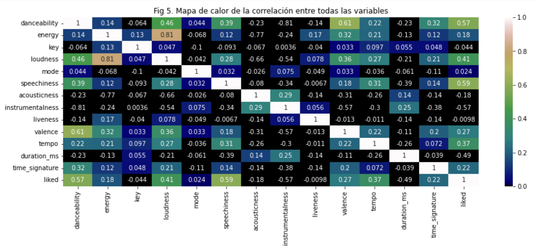
La tabla anterior es cuadrada y simétrica. Esto significa que la correlación que hay entre cualquier par de variables no está dictaminada por cuál de ambas es la variable dependiente y cuál de ellas la independiente.

Las correlaciones cuyo signo es positivo indican que ambas variables aumentan o decrementan su valor cuando la otra lo hace; es decir, su crecimiento se realiza hacia la misma dimensión.

En el caso contrario, la correlación negativa de las variables indica que el crecimiento de una variable provoca el decrecimiento de otra y viceversa.

In [38]:
```
# 5. Mapa de calor de todas las variables
plt.figure(figsize=(15,5))
sns.heatmap(songs.corr(), annot=True, vmin=0, vmax=1, cmap="gist_earth")

plt.title('Fig 5. Mapa de calor de la correlación entre todas las variables')
plt.show()
```



Al utilizar el mapa de calor podemos visualizar las correlaciones de las variables y cómo el comportamiento de una afecta a la otra.

Esta distribución de colores nos permite observar fácil y rápidamente que las celdas con valores oscuros representan valores más bajos que aquellas con colores claros. En este caso, *vim* es 0 y nos permite observar que las valores negativos están con color negro, lo cual facilita la búsqueda de valores de correlación positivos.

Un ejemplo de una correlación fuerte es la que existe entre *loudness* y *energy*, siendo el valor 0.81. Esto indica que cuando una canción tiene mucha energía es altamente probable que también tenga un volumen alto.

En contraste, se encuentran las variables *liked* e *instrumentalness*, pues la correlación es -0.57. Esto indica que cuando una canción tiene un valor alto de instrumentalización, la probabilidad de adquirir gusto por ella es muy baja.