In [1]:

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
from scipy import stats
sns.set(style="whitegrid")


%matplotlib inline
```

Warm UP

Read the data

In [2]:

```
users = pd.read_csv("user_data_sample.csv")
songs = pd.read_csv("end_song_sample.csv")
data = songs.merge(users, how='inner', on='user_id')
```

Users understanding

In [3]:

```
print('There is', len(users) ,'users registered')
users.head()
```

```
There is 9565 users registered
```

Out[3]:

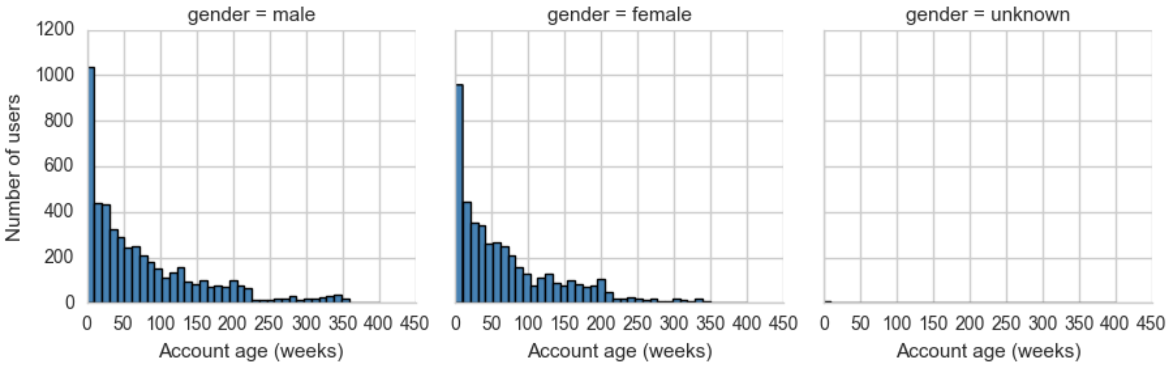
	gender	age_range	country	acct_age_weeks	user_id
0	male	25 - 29	FR	329	97f47c9fba714ca68320b8a80e010a1a
1	female	45 - 54	US	178	d615ca85849d458e9a5d755ec4727e8f
2	female	18 - 24	DE	68	6c83a5bf63b74f85b106ac7e7e015a1b
3	female	30 - 34	US	8	530fcedb3f244e6f91ecb326740005eb
4	female	30 - 34	FR	42	d2ed6a815eda4f61aa346b7936d03ef7

In [4]:

```
g = sns.FacetGrid(users, col="gender", margin_titles=True)
bins = np.linspace(0, 400, 40)
g.map(plt.hist, "acct_age_weeks", color="steelblue", bins=bins, lw=1).set_axis_labels("Acc
```

Out[4]:

```
<seaborn.axisgrid.FacetGrid at 0x1ce44b75470>
```

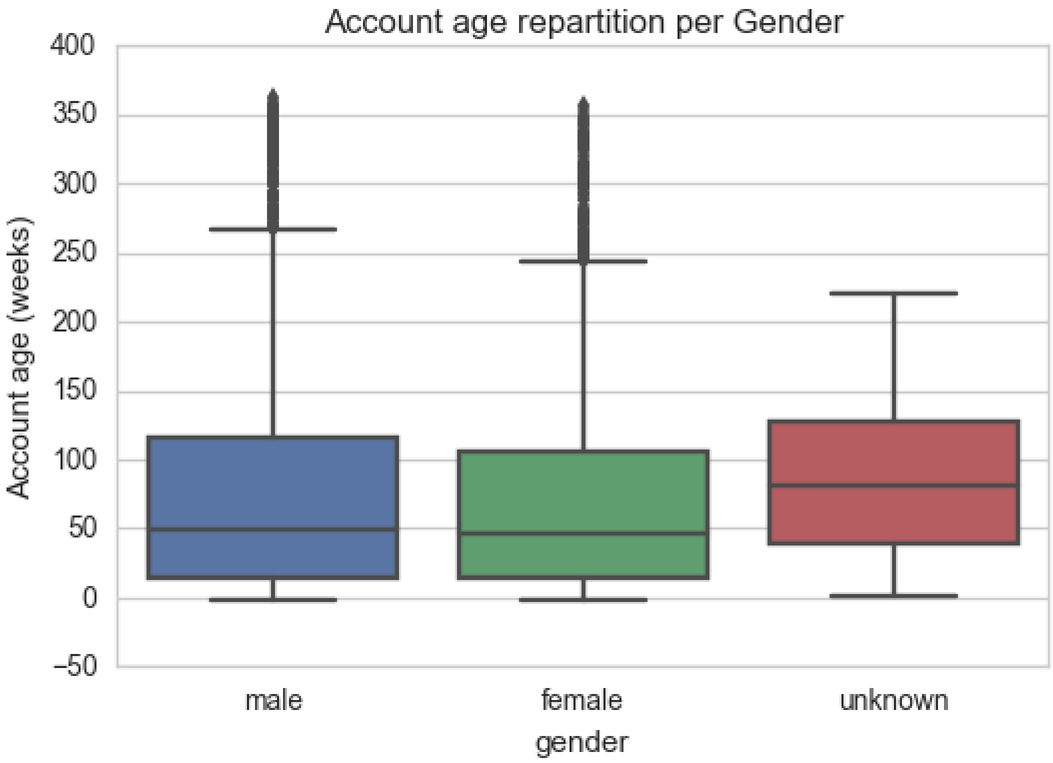


In [5]:

```
sns.boxplot(y="acct_age_weeks", x="gender", data=users)
plt.ylabel("Account age (weeks)")
plt.title("Account age repartition per Gender")
```

Out[5]:

```
<matplotlib.text.Text at 0x1ce45a51c18>
```



We see that the global distribution of the age of the user account looks pretty much the same for men and women. Most users are new on the producs for about **50 weeks**.

We also see that there is some users with **Unknown** gender. Let's see how many of them we have

In [6]:

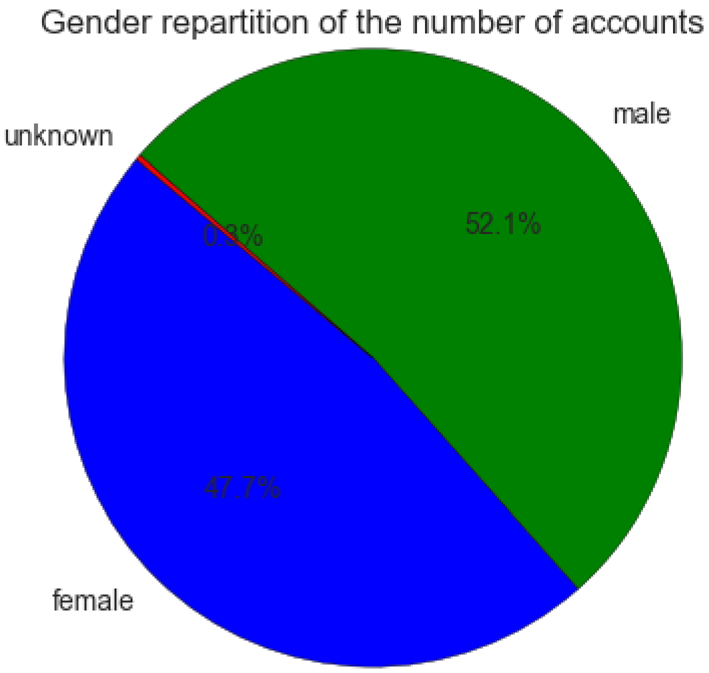
```
print('there is', len(users[users.gender == 'unknown']), 'users with unknown gender')
```

```
there is 26 users with unknown gender
```

They are not that many so we can see how to manage them later (maybe remove them)

In [7]:

```
gender = users.groupby(['gender'], as_index=False).agg({"user_id":pd.Series.nunique})
gender.columns = ['gender', 'count_users']
plt.title("Gender repartition of the number of accounts")
plt.pie(gender.count_users, labels=gender.gender,  autopct='%1.1f%%',startangle=140)
plt.axis('equal')

plt.show()
```

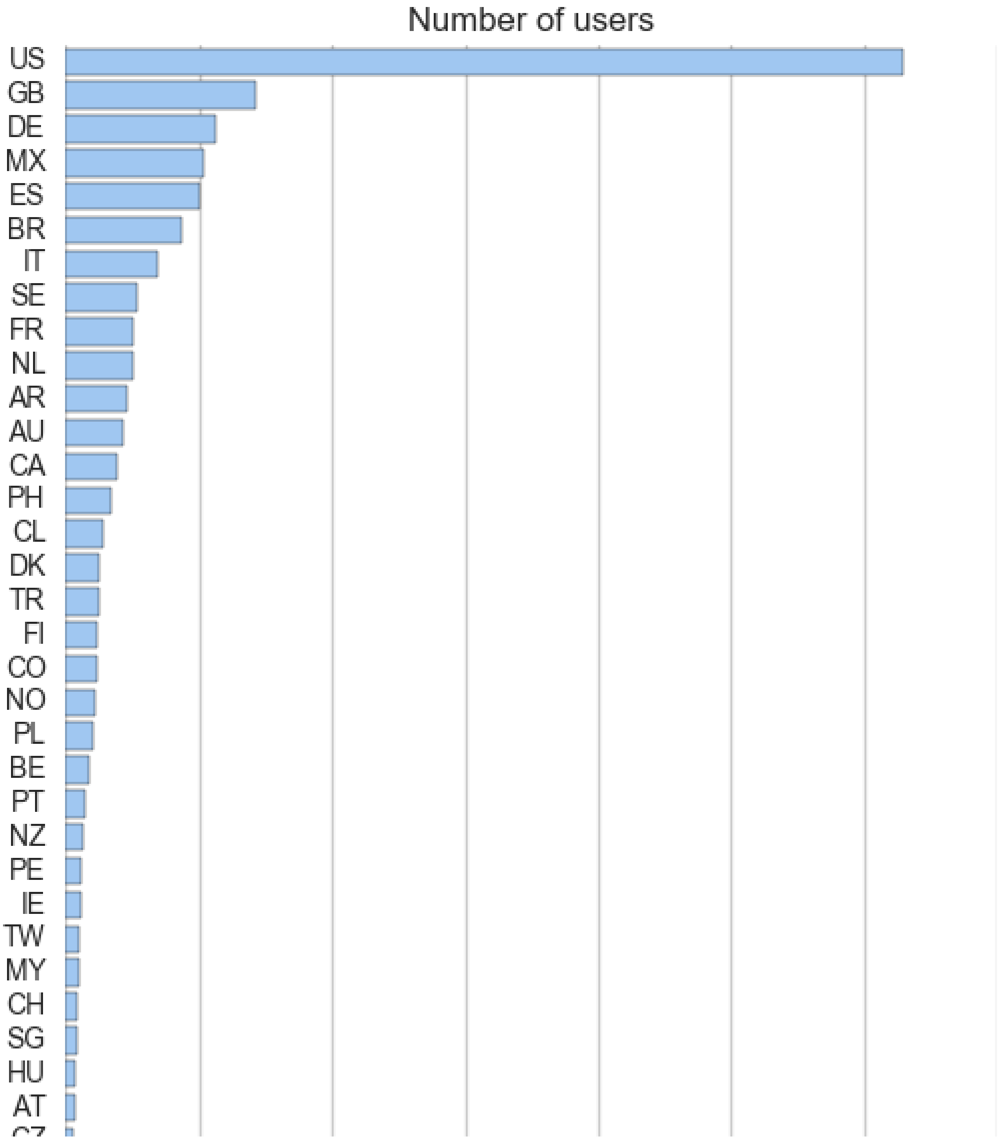


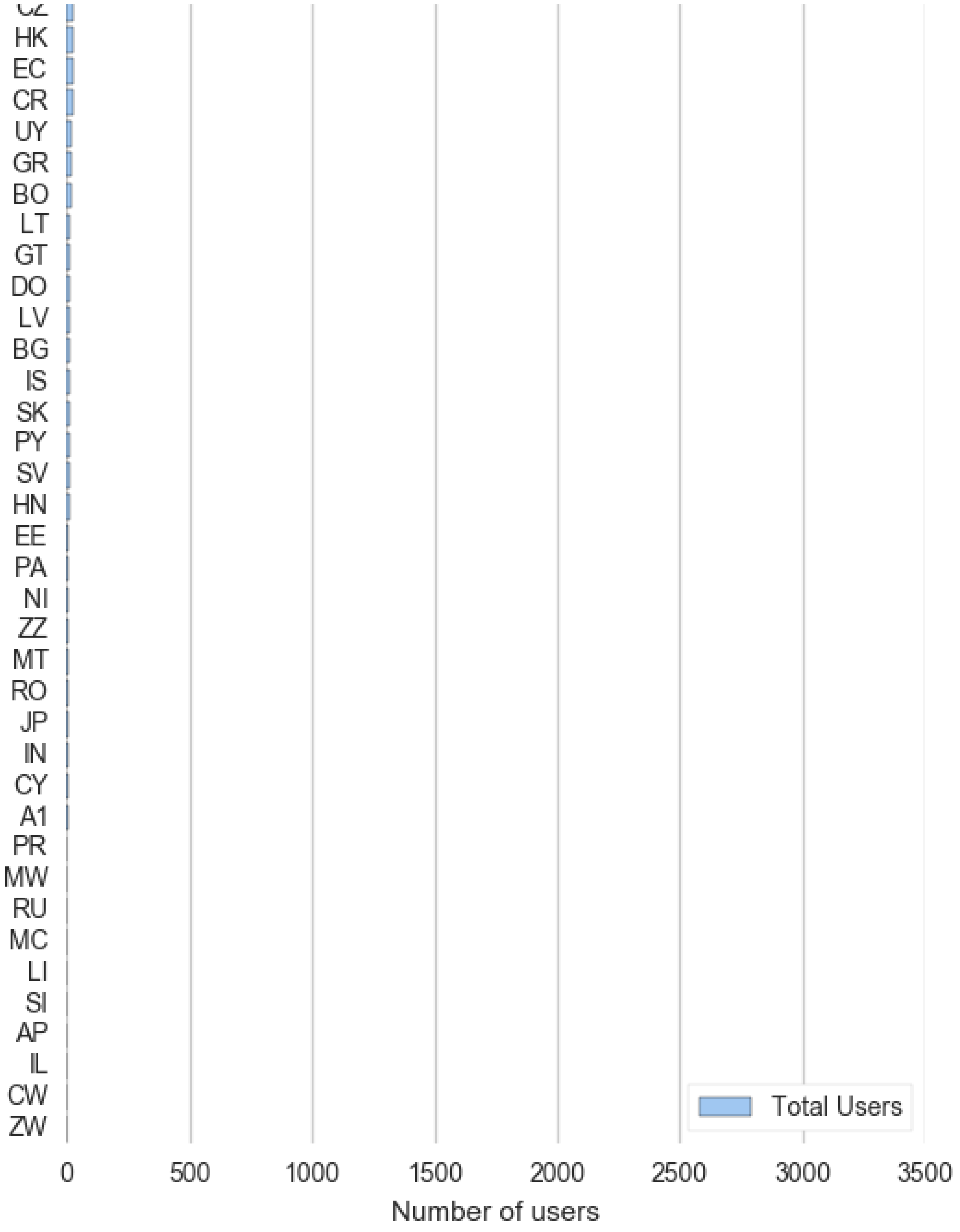
In [8]:

```
# Initialize the matplotlib figure
f, ax = plt.subplots(figsize=(6, 15))
country_gender = users.groupby(['country'], as_index=False).agg({"user_id":"count"})

country_gender.columns = ['country', 'count_users']

country_gender = country_gender.sort_values("count_users", ascending=False)

# Plot the all the users
sns.set_color_codes("pastel")
sns.barplot(x="count_users", y="country", data=country_gender,
            label="Total Users", color="b")


# Add a legend and informative axis label
ax.legend(ncol=2, loc="lower right", frameon=True)
ax.set(ylabel="",
       xlabel="Number of users")
sns.despine(left=True, bottom=True)
plt.title('Number of users')
#sns.barplot()
```

Out[8]:

```
<matplotlib.text.Text at 0x1ce45d19d30>
```



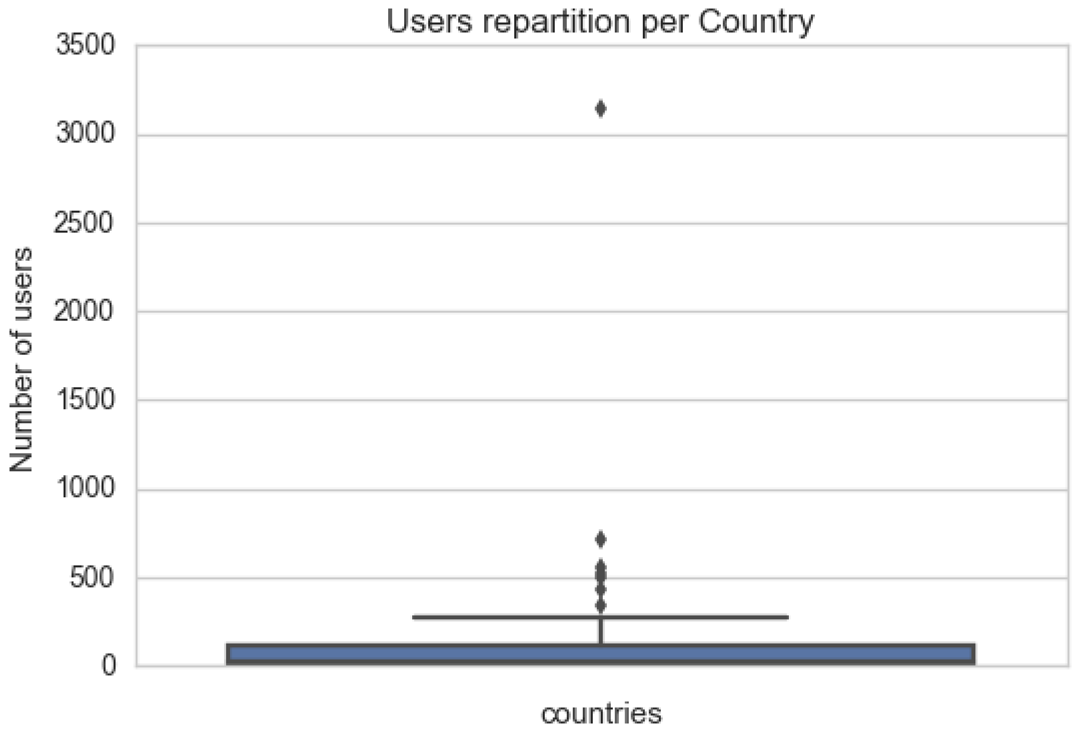

CZ
HK
EC
CR
UY
GR
BO
LT
GT
DO
LV
BG
IS
SK
PY
SV
HN
EE
PA
NI
ZZ
MT
RO
JP
IN
CY
A1
PR
MW
RU
MC
LI
SI
AP
IL
CW
ZW
0
500
1000
1500
2000
2500
3000
3500
Number of users
Total Users

In [16]:

```
sns.boxplot(country_gender.count_users, orient='v')
plt.xlabel('countries')
plt.ylabel('Number of users')
plt.title('Users repartition per Country')
```

Out[16]:

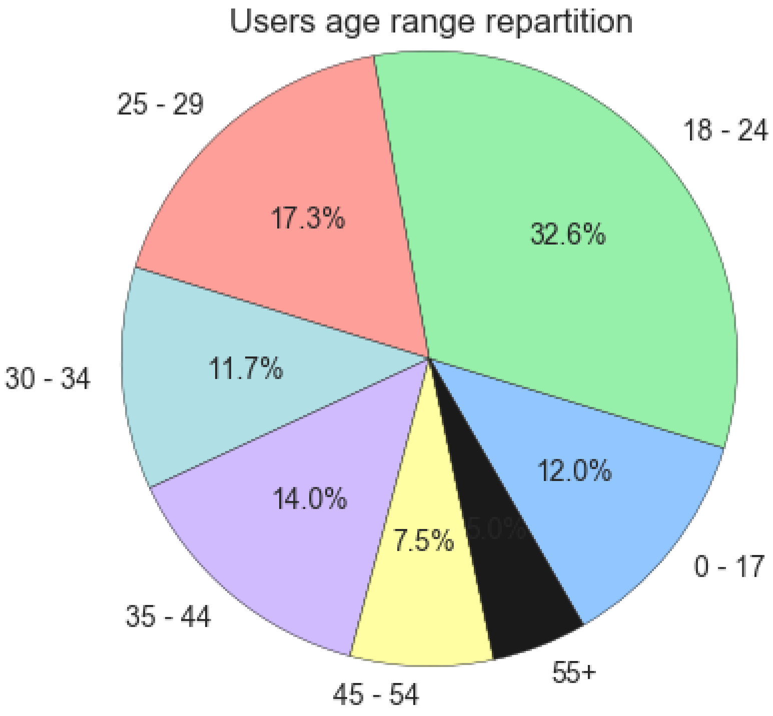
```
<matplotlib.text.Text at 0x1ce4652aba8>
```



The boxplot is very short which means that the countries have a more or less equivalent number of users

In [19]:

```
age = users.groupby(['age_range'], as_index=False).agg({"user_id":"count"})
age.columns = ['age_range', 'count_users']

plt.pie(age.count_users, labels=age.age_range,  autopct='%1.1f%%', startangle=300)
plt.axis('equal')
plt.title("Users age range repartition")
plt.show()
```



We see that beside the age range **18 - 24**, the rest of age range are equally distributed.

Tracks listening understanding

In [11]:

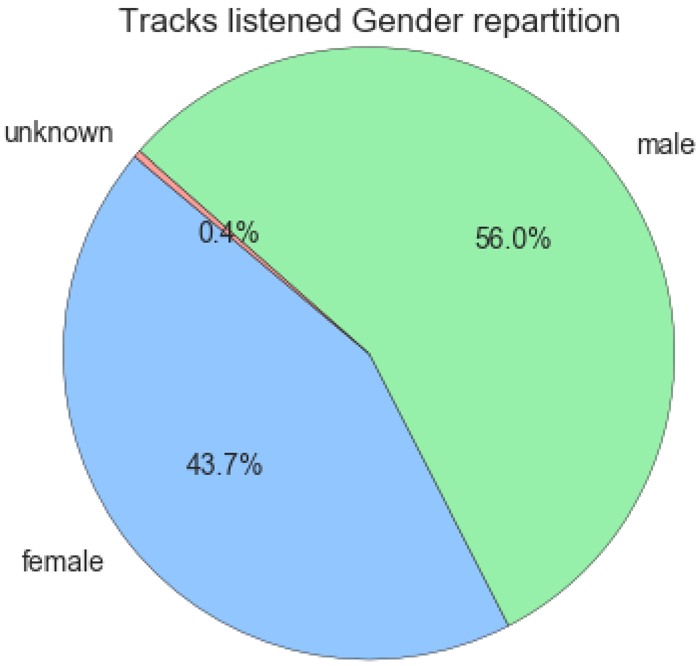
```
songs.describe()
```

Out[11]:

	ms_played	end_timestamp
count	1.342891e+06	1.342891e+06
mean	1.287120e+05	1.444270e+09
std	1.200548e+05	3.518090e+05
min	0.000000e+00	1.443658e+09
25%	3.778000e+03	1.443964e+09
50%	1.476780e+05	1.444272e+09
75%	2.228010e+05	1.444574e+09
max	5.100017e+06	1.444867e+09

In [12]:

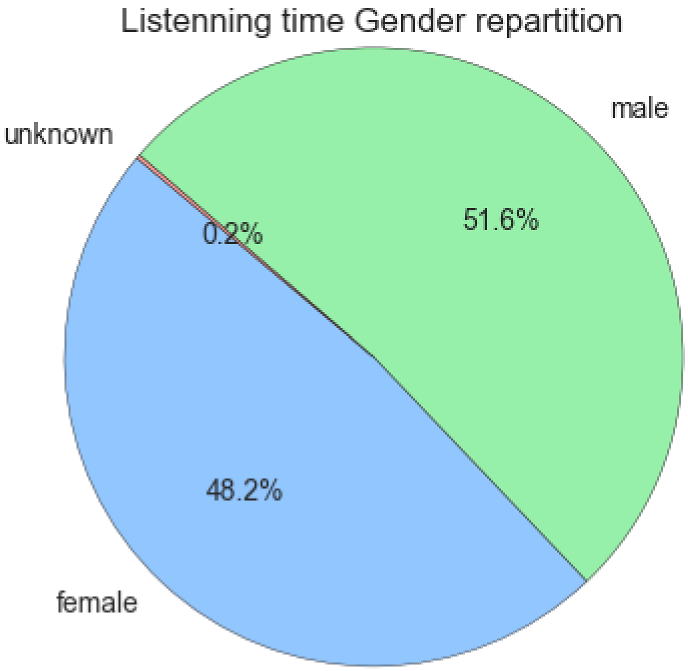
```
count_tracks_per_gender = data.groupby(['gender'], as_index=False).agg({"track_id":pd.Serie
count_tracks_per_gender.columns = ['gender', 'count_tracks']
plt.title("Tracks listened Gender repartition")
plt.pie(count_tracks_per_gender.count_tracks, labels=count_tracks_per_gender.gender,
    autopct='%1.1f%%',startangle=140)
plt.axis('equal')

plt.show()
```



The reparition of differentes tracks listenened is slightly different than the repartition of the users. Are men slightly more diverse in their taste of music than women?? (We'll confirm that hypothesis with a statistical test)

In [13]:

```
listen_time_per_gender = data.groupby(['gender'], as_index=False).agg({"ms_played":"sum"})
listen_time_per_gender.columns = ['gender', 'sum_listen_time']
plt.title("Listenning time Gender repartition")
plt.pie(listen_time_per_gender.sum_listen_time, labels=listen_time_per_gender.gender,
    autopct='%1.1f%%',startangle=140)
plt.axis('equal')

plt.show()
```



At first sight, We can assume that male and female listeners are pretty much the same in their overall listening.
Let's confirm it via some statisticals tests

Statisticals Tests

Tracks diversity

Let's see if men are as diverse as women in term of tracks listened.

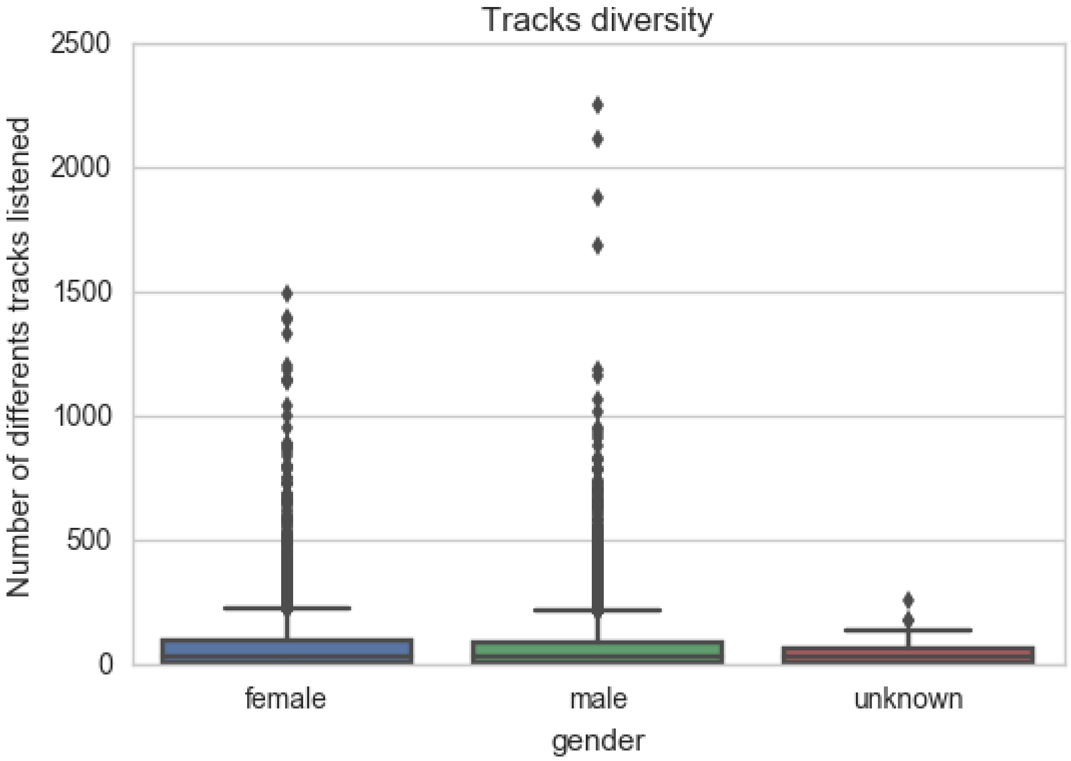
In [14]:

```
count_tracks = data.groupby(['gender', 'user_id'], as_index=False).agg({"track_id":pd.Serie
count_tracks.columns = ['gender', 'user_id', 'count_tracks']

count_tracks_male = count_tracks[count_tracks.gender == 'male']
count_tracks_female = count_tracks[count_tracks.gender == 'female']

print("Student Test")
print("Null Hypothesis for the test : The distribution of the number of tracks listened is
t, pvalue = stats.ttest_ind(count_tracks_male.count_tracks, count_tracks_female.count_track
print("p-value = {0:.3f}".format(pvalue))

sns.boxplot(data=count_tracks, x="gender", y="count_tracks")
plt.ylabel('Number of differents tracks listened')
plt.title('Tracks diversity')
```

```
Student Test
Null Hypothesis for the test : The distribution of the number of tracks list
ened is the same for men and women
p-value = 0.955
```

Out[14]:

```
<matplotlib.text.Text at 0x1ce4646d1d0>
```



With a p-value of **0.955** (which is far greater than *0.05*) we can assume that the null hypothesis is right.
So, **The gender has no influence on the tracks diversity**

Listening time

In terms of the count of listening time, let's test if male and female listeners are significantly different in their overall listening

In [15]:

```
listen_time = data.groupby(['gender', 'user_id'], as_index=False).agg({"ms_played":"sum"})
listen_time.columns = ['gender','user_id' ,'sum_listen_time']

listen_time_male = listen_time[listen_time.gender == 'male']
listen_time_female = listen_time[listen_time.gender == 'female']

print("Student Test")
print("Null Hypothesis for the test : The distribution of the listening time is the same fo
t, pvalue = stats.ttest_ind(listen_time_male.sum_listen_time, listen_time_female.sum_listen
print("p-value = {0:.3f}".format(pvalue))
```

```
Student Test
Null Hypothesis for the test : The distribution of the listening time is the
 same for men and women
p-value = 0.621
```

With a p-value of **0.621** (which is far greater than *0.05*) we can assume that the null hypothesis is right.
So, **The gender has no influence on the Listening time**

Check the next part **1.Session Breakdown**