In [1]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import calendar
import datetime
import matplotlib.dates as mdates
sns.set_theme(style="whitegrid")

d = dict((v,k) for k,v in enumerate(calendar.month_name))
from statsmodels.stats.proportion import proportion_confint

# Helper function
def f_gen(column):
    def f(x):
        d = {}
        N_not_return = sum(x['not_returned'])
        N_total = sum(x['total_checkouts'])
        d[column] = x[column].values[0] + "(N = {})".format(N_total)
        d['month_val'] = x['month_val'].values[0]
        d['Non-Return Rate'] = N_not_return/N_total
        d['N_not_returned'] = N_not_return
        d['N_total'] = N_total
        lb, ub = proportion_confint(N_not_return, N_total)
        d['ci_lb'] = lb
        d['ci_ub'] = ub
        return pd.Series(d)
    return f
```

Concept

I wanted to explore the items in the SPL that were checked out and never returned. More specifically, I was interested in understanding what types are most likely to be not returned and whether this changes over time. This study uses all of the transactions between 2016 and 2019. I would have preferred to use all of the data, but larger queries took too long to provide results.

Query

```
SELECT
    i.itemType,
    z.month,
    z.year,
    SUM(z.not_returned) AS not_returned,
    SUM(z.total_checkouts) AS total_checkouts
FROM
    (SELECT
        MONTHNAME(t.checkOut) AS month,
            YEAR(t.checkOut) AS year,
            DAY(t.checkOut) AS day,
            t.itemNumber,
            COUNT(t.checkOut) AS checkouts,
            COUNT(t.checkIn) AS checkins,
            CASE
                WHEN COUNT(t.checkOut) / 2 - COUNT(t.checkIn) > 0 THEN 1
                ELSE 0
            END AS not_returned,
            CASE
                WHEN COUNT(t.checkOut) - COUNT(t.checkIn) > 0 THEN COUNT(t.checkOut) -
COUNT(t.checkIn)
                ELSE COUNT(t.checkOut)
            END AS total_checkouts
    FROM
        spl_2016.transactions t
    WHERE
```

```
        t.checkOut > DATE(NOW() - INTERVAL 5 YEAR)
            AND t.checkOut < DATE(NOW() - INTERVAL 1 YEAR)
    GROUP BY t.checkOut , t.itemNumber) z
        INNER JOIN
    spl_2016.itemType i ON z.itemNumber = i.itemNumber
GROUP BY i.itemType , z.month , z.year
```

Query Design

The above query uses the transactions to find all of the check-outs that do not have a respective check-in. One challenge in this is that the transaction table records two entries for every transaction - one when checking out for the first time, and one when checking in. This prevents us from doing a simple select where check in is null because this only selects the first checkout times and doesn't take into account whether a subsequent check in is done.

A heuristic to use to account for this without having to do a large join is, if we group by the checkOut and itemNumber, then if there are 2 checkOut and 1 checkIn, then we can safely say the item has been returned. Alternatively, if there is only 1 checkOut and 0 checkIns, then we can say that the item was never returned. This works for the majority of cases but there are two anomalous edge cases that I found during the exploration of this.

1. It is possible that, instead of 2 rows for the transaction, there is only 1 where there is checkOut and checkIn is filled in. Likely, this is due to a missing entry for a checkOut only event, but regardless, this is counted as a returned item.

2. Sometimes, there are > 2 checkout events for a single checkOut, itemNumber pair. I'm not entirely sure why this is but in early investigations, it seems that most often, it's 3 checkOuts and 2 checkIns, which would seem to place it as a successful return. This is built into the above logic.

Once we are able to mark the transactions that were never returned, we can group them however we want to investigate what types of items tend to not be returned. I chose to group by itemType for this case because it is a manageable and well-documented field for each transaction.

Alternative Query

An alternative query would be to grab all of the entries in the outraw table that do not have a respective entry in the inraw table but this would require an expensive join operation and results in much worse performance. A quick double check verifies that the transaction table is simply the inraw and the outraw table combined.

Plotting and Interpretation

To analyze, this data, I first investigated some basic statistics about non-return rates.

In [2]:

```
data = pd.read_csv("spl_data.csv")
item_type_ref = pd.read_csv("integrated-library-system-ils-data-dictionary.csv")
data['month_val'] = data['month'].map(d)
data = pd.merge(data, item_type_ref, left_on = 'itemType', right_on = 'Code')
data = data[data['Format Subgroup'].notna()]
```

Looking at the data in aggregate, we see that more than a quarter of items are not returned. This is a surprisingly large number of items.

In [3]:

```
print("Aggregate Non-Return Rate: {:.2f}%".format(sum(data['not_returned']) * 100/sum(data['total_ch
```

```
Aggregate Non-Return Rate: 26.67%
```

Non-Return Rates by Media Type

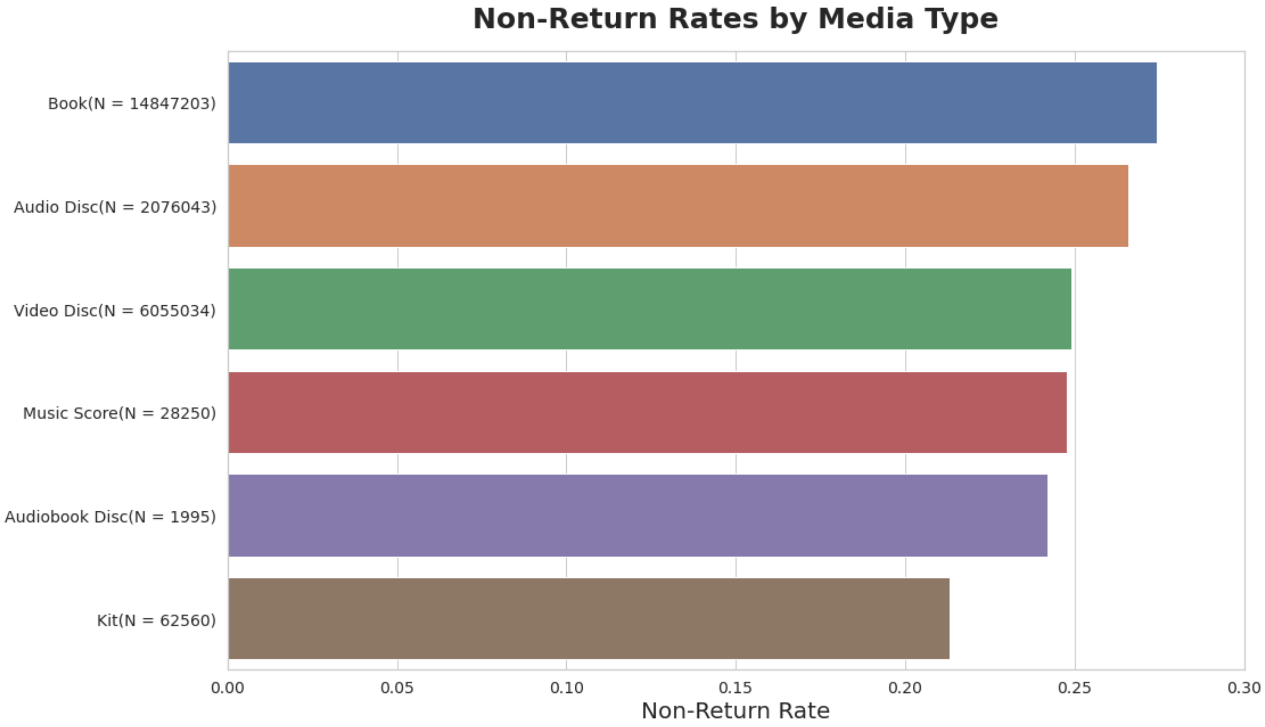
Breaking it down into the type of media, we see that books tend to lead, followed by audio, and then video. Music scores are interesting as well, although the system in which those are checked out and not returned is mysterious to me.

In [4]:

```
df = data.groupby(['Format Subgroup']).apply(f_gen("Format Subgroup")).sort_values('Non-Return Rate'
df = df[df['N_total'] > 200]

fig, ax = plt.subplots()
sns.barplot(y = "Format Subgroup", x = "Non-Return Rate", data = df, ci = 0, ax = ax)
ax.set_title("Non-Return Rates by Media Type", fontsize = 25, fontweight = 'bold', pad = 20)
ax.set_xlabel("Non-Return Rate", fontsize = 20)
ax.set_ylabel("", fontsize = 20)
ax.tick_params(labelsize = 14)

ax.set_xlim([0,0.30])
fig.set_size_inches(16, 10)
```



In [5]:

```
df
```

Out[5]:

	Format Subgroup	month_val	Non-Return Rate	N_not_returned	N_total	ci_lb	ci_ub
Format Subgroup							
Book	Book(N = 14847203)	4	0.274361	4073495	14847203	0.274134	0.274588
Audio Disc	Audio Disc(N = 2076043)	4	0.265859	551935	2076043	0.265258	0.266460
Video Disc	Video Disc(N = 6055034)	4	0.248992	1507653	6055034	0.248647	0.249336
Music Score	Music Score(N = 28250)	4	0.247752	6999	28250	0.242718	0.252786
Audiobook Disc	Audiobook Disc(N = 1995)	4	0.242105	483	1995	0.223308	0.260902
Kit	Kit(N = 62560)	4	0.213187	13337	62560	0.209978	0.216397

Non-Return Rates by Book Type (Fiction vs Non-Fiction vs Reference)

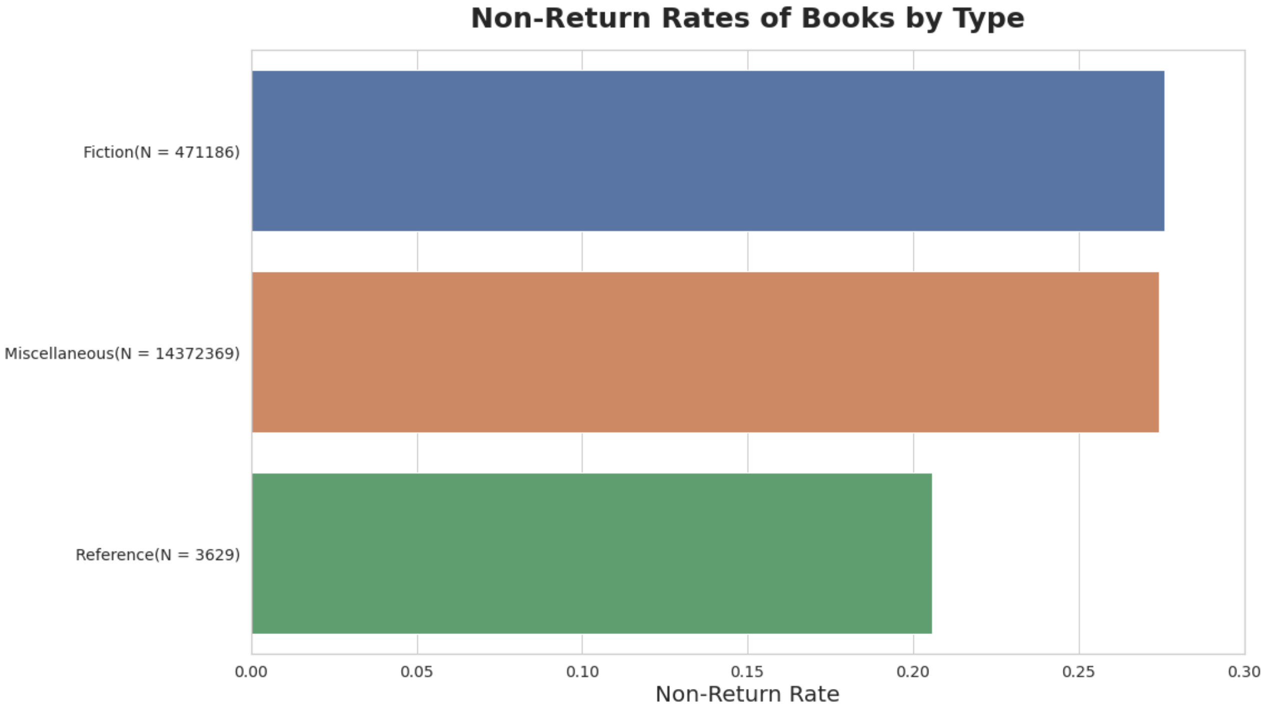
As books form the largest category of non-returned items, it is also interesting to look at the comparison of fiction vs reference vs others. From this we can see that there is no significant difference between fiction and other items but reference books are significantly lower than either of them. The fact that reference items are non-zero is surprising in itself as those are a type of book that should always be returned.

In [6]:

```
df = data[data['Format Subgroup'] == 'Book'].groupby('Category Group').apply(f_gen('Category Group')
df = df[df['N_total'] > 200]

fig, ax = plt.subplots()
sns.barplot(y = "Category Group", x = "Non-Return Rate", data = df, ci = 0, ax = ax)
ax.set_title("Non-Return Rates of Books by Type", fontsize = 25, fontweight = 'bold', pad = 20)
ax.set_xlabel("Non-Return Rate", fontsize = 20)
ax.set_ylabel("", fontsize = 20)
ax.tick_params(labelsize = 14)

ax.set_xlim([0,0.30])
fig.set_size_inches(16, 10)
```



In [7]:

```
df
```

Out[7]:

	Category Group	month_val	Non-Return Rate	N_not_returned	N_total	ci_lb	ci_ub
Category Group							
Fiction	Fiction(N = 471186)	4	0.275821	129963	471186	0.274545	0.277097
Miscellaneous	Miscellaneous(N = 14372369)	4	0.274331	3942782	14372369	0.274100	0.274561
Reference	Reference(N = 3629)	4	0.205842	747	3629	0.192687	0.218996

Non-Return Rates over Time By Media Type

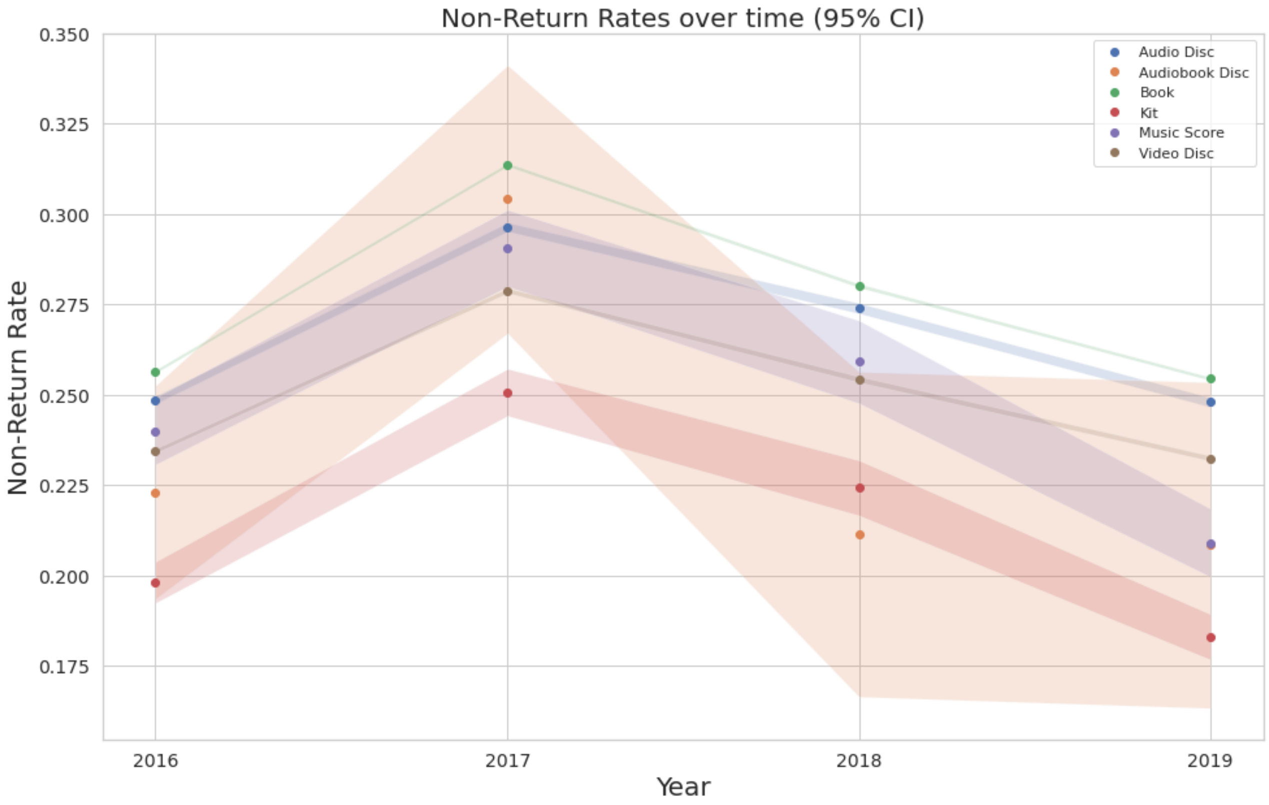
Another dimension to look at the same data is how the return rates changed over time. We can see that in 2017, there was a significant spike in non-return rates over all categories.

In [8]:

```
df = data.groupby(['Format Subgroup', 'year']).apply(f_gen("Format Subgroup"))
fig, ax = plt.subplots()
for media_type, v in df.groupby(level=0):
    if sum(v['N_total']) > 200:
        years = [2016,2017,2018,2019]
        nrr = []
        lb_nrr = []
        ub_nrr = []
        for y in years:
            nrr.append(v.loc[(media_type, y)]['Non-Return Rate'])
            lb_nrr.append(v.loc[(media_type, y)]['ci_lb'])
            ub_nrr.append(v.loc[(media_type, y)]['ci_ub'])
        ax.plot_date([datetime.datetime.strptime(str(y), '%Y') for y in years], nrr, label = media_t
        ax.fill_between([datetime.datetime.strptime(str(y), '%Y') for y in years], lb_nrr, ub_nrr, a

years = mdates.YearLocator()   # every year
years_fmt = mdates.DateFormatter('%Y')

# format the ticks
ax.xaxis.set_major_locator(years)
ax.xaxis.set_major_formatter(years_fmt)
ax.set_xlabel("Year", fontsize = 20)
ax.set_ylabel("Non-Return Rate", fontsize = 20)
ax.set_title("Non-Return Rates over time (95% CI)", fontsize = 20)
ax.tick_params(labelsize = 14)
ax.legend()
fig.set_size_inches(16, 10)
```



Non-Return Rates over Time By Month

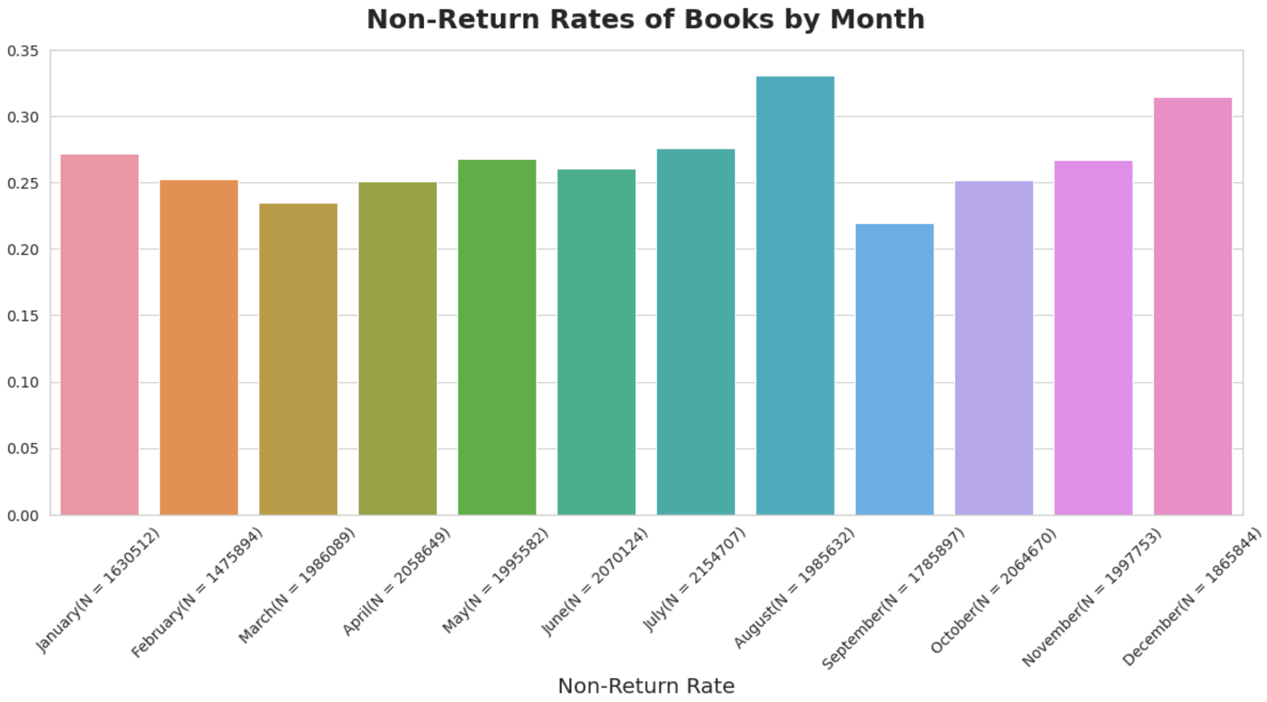
Another dimension to look at the same data is how the return rates change per month. It seems like August has the highest rate of non-return followed by December. The reason for this is not so clear - December could be an end of the year effect.

In [9]:

```
df = data.groupby('month').apply(f_gen('month')).sort_values('month_val', ascending = True)
df = df[df['N_total'] > 200]

fig, ax = plt.subplots()
sns.barplot(x = "month", y = "Non-Return Rate", data = df, ax = ax)
ax.set_title("Non-Return Rates of Books by Month", fontsize = 25, fontweight = 'bold', pad = 20)
ax.set_xlabel("Non-Return Rate", fontsize = 20)
ax.set_ylabel("", fontsize = 20)
ax.tick_params(labelsize = 14)
ax.tick_params(axis = "x", rotation = 45)

ax.set_ylim([0,0.35])
fig.set_size_inches(20, 8)
```



Analysis

Surprisingly, according to the query, between 2016 and 2020, approximately 25% of items checked out were never returned. Breaking down by item type, books are consistently the most borrowed but not returned items, followed by audio and then video. This ordering may be justified by the fact that books tend to take a longer time to go through and thus is something that a patron can simply borrow and never get to, eventually not returning it.

Breaking down return rate by book type, it seems that there really is no preference for keeping fiction books and non-fiction (miscellaneous) books. Reference books fall much lower but that's understandable as something that is not supposed to be borrowed. The fact that reference books have outstanding copies brings into question the interpretation a bit for this query as we would expect reference books to be always be returned.

If we look at the non-return rates over time by item type, we see an interesting trend where it seems that it 2017, non-return rates spiked in all categories but since then have been trending downards proportionally. The reason for this is difficult to determine. Possiblilities include policy changes in the library, if we believe that the query returns the correct numbers and interpretation, or simply a change in the system to reflect a more efficient and accurate system.

Several item types were not included in these plots because they fell below the threshold of having at least 200 total checkout events.