EDA_Task-1_Maithil

November 14, 2022

[2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[3]:
```
data = pd.read_csv('train.csv')
data.head()
```

[3]:
```
   row_id        date  country       store                              product  \
0       0  01-01-2017  Belgium  KaggleMart        Kaggle Advanced Techniques
1       1  01-01-2017  Belgium  KaggleMart            Kaggle Getting Started
2       2  01-01-2017  Belgium  KaggleMart                Kaggle Recipe Book
3       3  01-01-2017  Belgium  KaggleMart  Kaggle for Kids: One Smart Goose
4       4  01-01-2017  Belgium  KaggleRama        Kaggle Advanced Techniques

   num_sold
0       663
1       615
2       480
3       710
4       240
```

[4]:
```
data.isnull().sum()
```

[4]:
```
row_id      0
date        0
country     0
store       0
product     0
num_sold    0
dtype: int64
```

[5]:
```
data = data.drop('row_id', axis=1)
```

[6]:
```
data.head()
```

[6]:

	date	country	store	product	num_sold
0	01-01-2017	Belgium	KaggleMart	Kaggle Advanced Techniques	663
1	01-01-2017	Belgium	KaggleMart	Kaggle Getting Started	615
2	01-01-2017	Belgium	KaggleMart	Kaggle Recipe Book	480
3	01-01-2017	Belgium	KaggleMart	Kaggle for Kids: One Smart Goose	710
4	01-01-2017	Belgium	KaggleRama	Kaggle Advanced Techniques	240

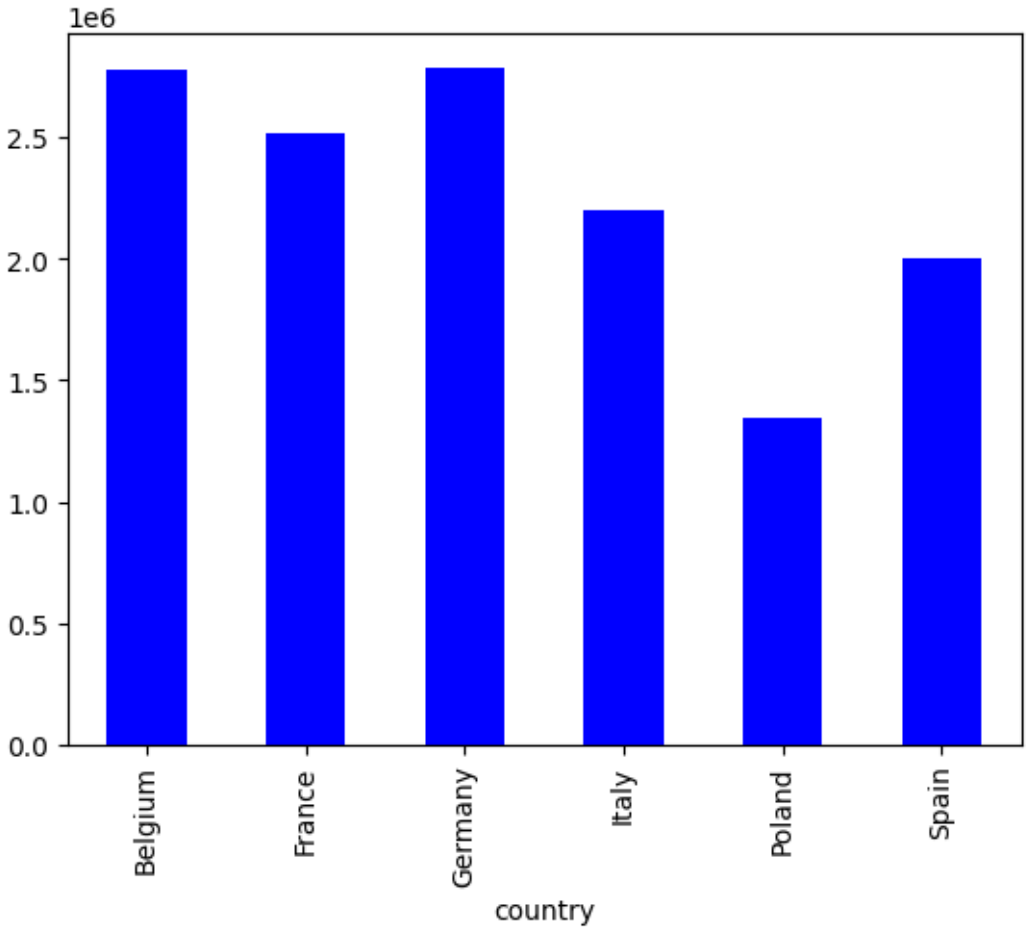
[7]:
```
data.date.value_counts()
```

[7]:
```
01-01-2017    48
10-09-2019    48
08-09-2019    48
07-09-2019    48
06-09-2019    48
              ..
01-05-2018    48
30-04-2018    48
29-04-2018    48
28-04-2018    48
31-12-2020    48
Name: date, Length: 1461, dtype: int64
```

[8]:
```
data1 = data.copy()
```

[10]:
```
sum_target = data1.groupby('country')['num_sold'].sum()
sum_target.plot(kind='bar', color='blue')
```

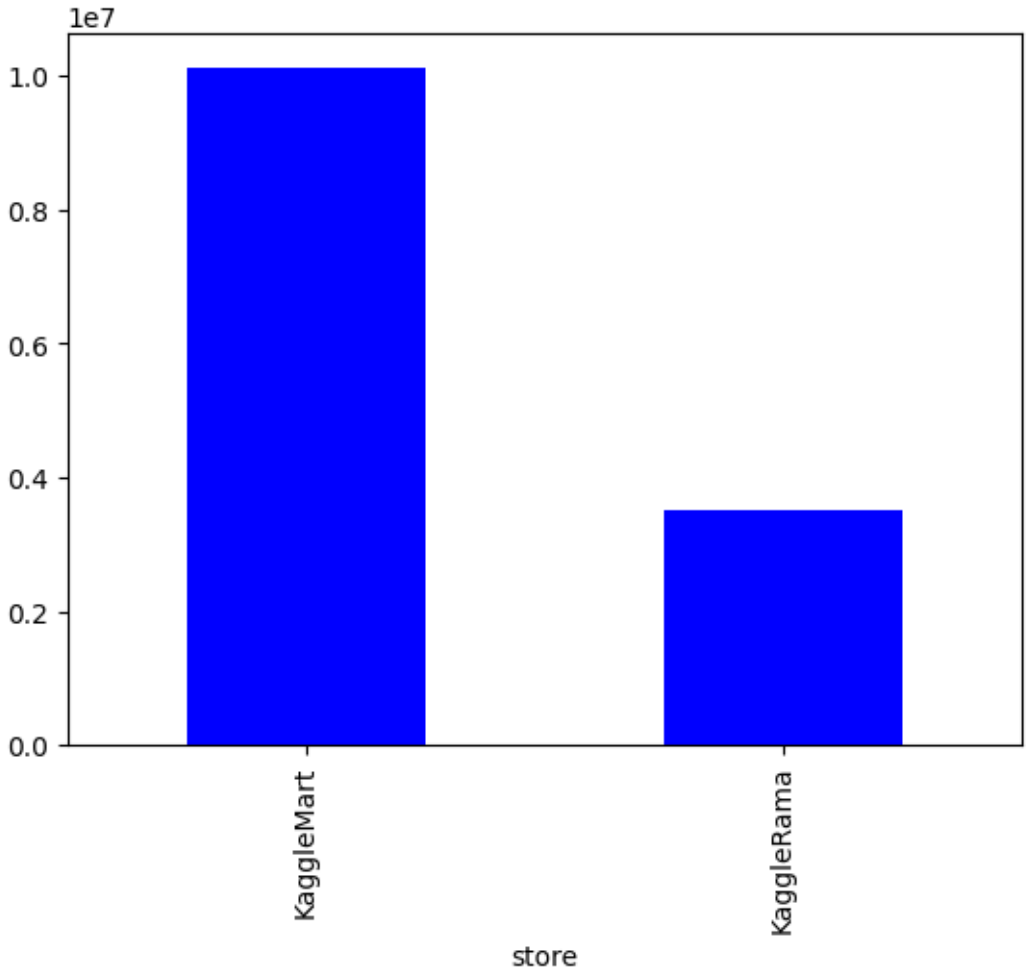
[10]:
```
<AxesSubplot: xlabel='country'>
```



The highest number of products are sold in Germany and Belgium

[11]:
```
sum2 = data1.groupby('store')['num_sold'].sum()
sum2.plot(kind='bar', color='blue')
```

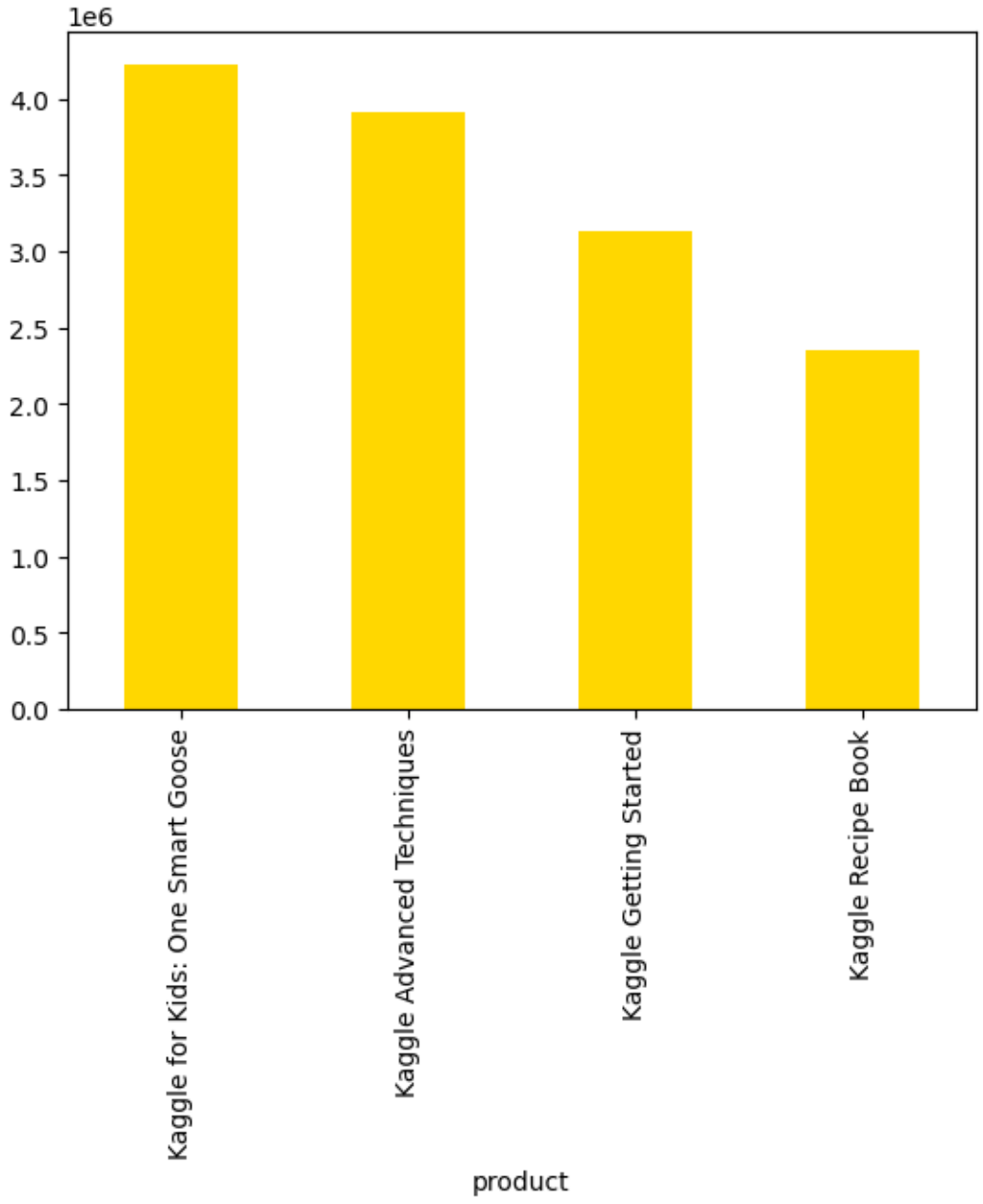
[11]:
```
<AxesSubplot: xlabel='store'>
```



The Kaggle Mart sold more amount of products than Kaggle Rama

[16]:
```
sum3 = data1.groupby('product')['num_sold'].sum()
sum3.sort_values(ascending=False).head(10).plot(kind='bar', color='gold')
```

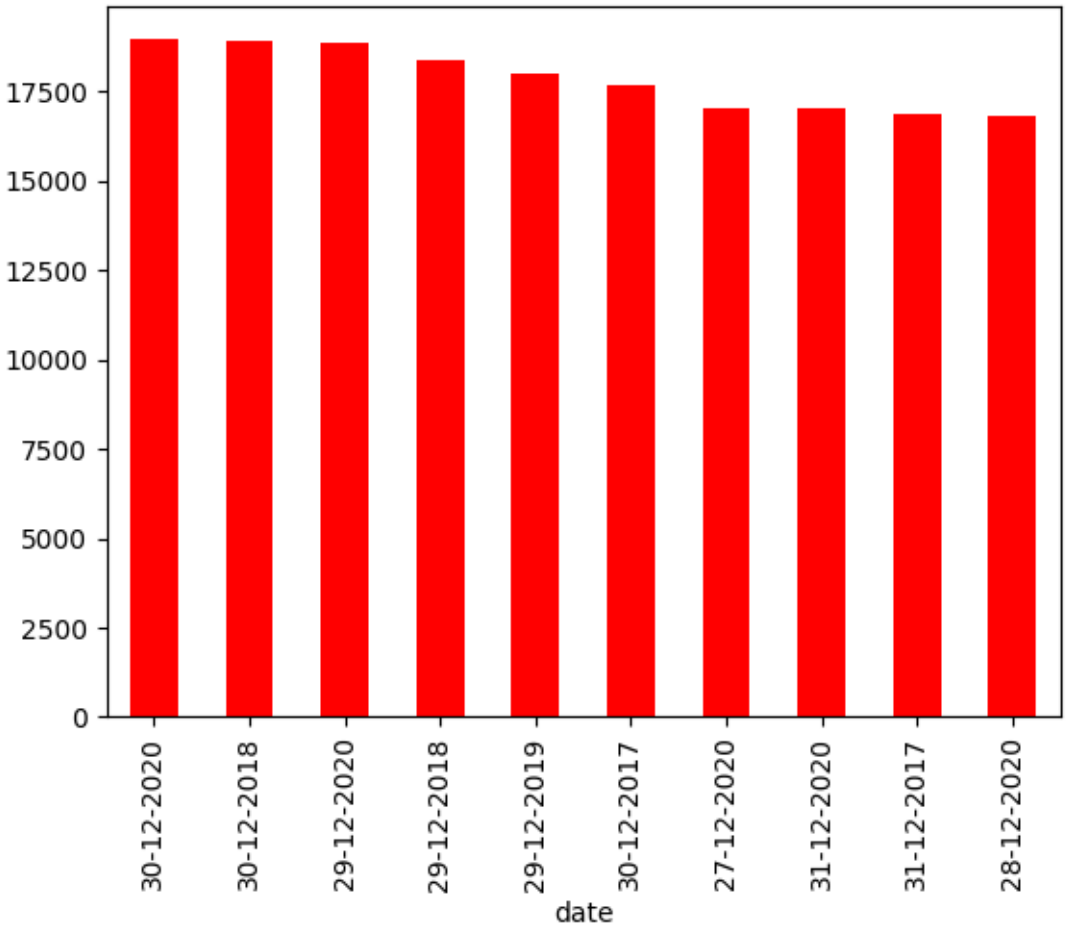
[16]:
```
<AxesSubplot: xlabel='product'>
```



The product which is sold the most is "Kaggle for Kids: One Smart Goose"

[17]:
```
sum4 = data1.groupby('date')['num_sold'].sum()
sum4.sort_values(ascending=False).head(10).plot(kind='bar', color='red')
```

[17]: <AxesSubplot: xlabel='date'>



[18]:
```
data.head()
```

[18]:

	date	country	store	product	num_sold
0	01-01-2017	Belgium	KaggleMart	Kaggle Advanced Techniques	663
1	01-01-2017	Belgium	KaggleMart	Kaggle Getting Started	615
2	01-01-2017	Belgium	KaggleMart	Kaggle Recipe Book	480
3	01-01-2017	Belgium	KaggleMart	Kaggle for Kids: One Smart Goose	710
4	01-01-2017	Belgium	KaggleRama	Kaggle Advanced Techniques	240

These are the top 10 dates when the products were sold most in numbers

[19]:
```
data.date = pd.to_datetime(data.date)
data.info()
data.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70128 entries, 0 to 70127
Data columns (total 5 columns):
```

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   date      70128 non-null  datetime64[ns]
 1   country   70128 non-null  object
 2   store     70128 non-null  object
 3   product   70128 non-null  object
 4   num_sold  70128 non-null  int64
dtypes: datetime64[ns](1), int64(1), object(3)
memory usage: 2.7+ MB
```

```
C:\Users\7520\AppData\Local\Temp\ipykernel_15356\1253929069.py:2: UserWarning:
Parsing dates in DD/MM/YYYY format when dayfirst=False (the default) was
specified. This may lead to inconsistently parsed dates! Specify a format to
ensure consistent parsing.
  data.date = pd.to_datetime(data.date)
```

[19]:

	date	country	store	product	num_sold
0	2017-01-01	Belgium	KaggleMart	Kaggle Advanced Techniques	663
1	2017-01-01	Belgium	KaggleMart	Kaggle Getting Started	615
2	2017-01-01	Belgium	KaggleMart	Kaggle Recipe Book	480
3	2017-01-01	Belgium	KaggleMart	Kaggle for Kids: One Smart Goose	710
4	2017-01-01	Belgium	KaggleRama	Kaggle Advanced Techniques	240

[47]:
```
data['Month'] = data.date.dt.month_name()
data['Date'] = data.date.dt.day
data['Day'] = data.date.dt.day_name()
data['Year'] = data.date.dt.year
data.head()
```

[47]:

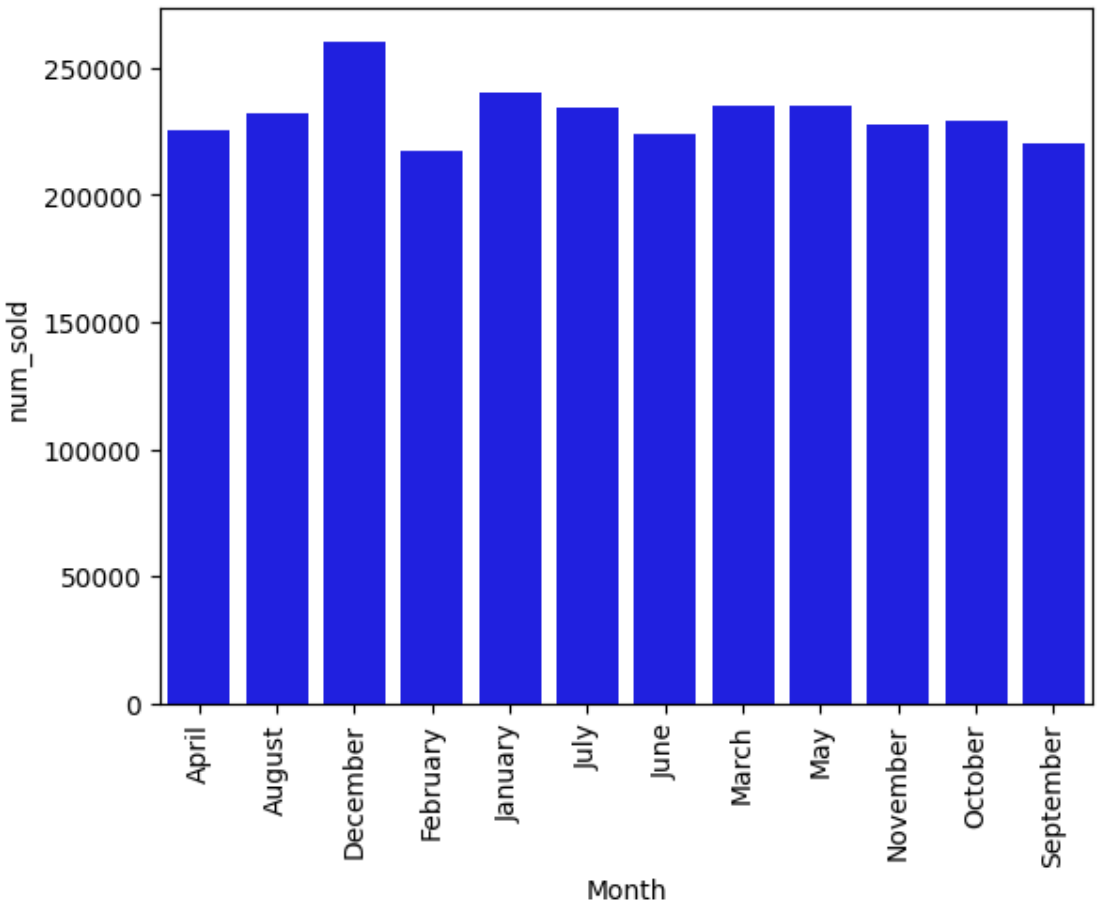
	date	country	store	product	num_sold	\
0	2017-01-01	Belgium	KaggleMart	Kaggle Advanced Techniques	663	
1	2017-01-01	Belgium	KaggleMart	Kaggle Getting Started	615	
2	2017-01-01	Belgium	KaggleMart	Kaggle Recipe Book	480	
3	2017-01-01	Belgium	KaggleMart	Kaggle for Kids: One Smart Goose	710	
4	2017-01-01	Belgium	KaggleRama	Kaggle Advanced Techniques	240	

	Month	Date	Day	Year
0	January	1	Sunday	2017
1	January	1	Sunday	2017
2	January	1	Sunday	2017
3	January	1	Sunday	2017
4	January	1	Sunday	2017

0.1 Belgium

[48]:
```
Belgium = data[data.country == 'Belgium']
Belgium = pd.DataFrame(Belgium.groupby('Month')['num_sold'].sum())
Belgium.reset_index(inplace=True)
sns.barplot(x='Month', y='num_sold', data=Belgium, color='blue')
plt.xticks(rotation=90)
```

[48]:
```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11]),
 [Text(0, 0, 'April'),
  Text(1, 0, 'August'),
  Text(2, 0, 'December'),
  Text(3, 0, 'February'),
  Text(4, 0, 'January'),
  Text(5, 0, 'July'),
  Text(6, 0, 'June'),
  Text(7, 0, 'March'),
  Text(8, 0, 'May'),
  Text(9, 0, 'November'),
  Text(10, 0, 'October'),
  Text(11, 0, 'September')])
```

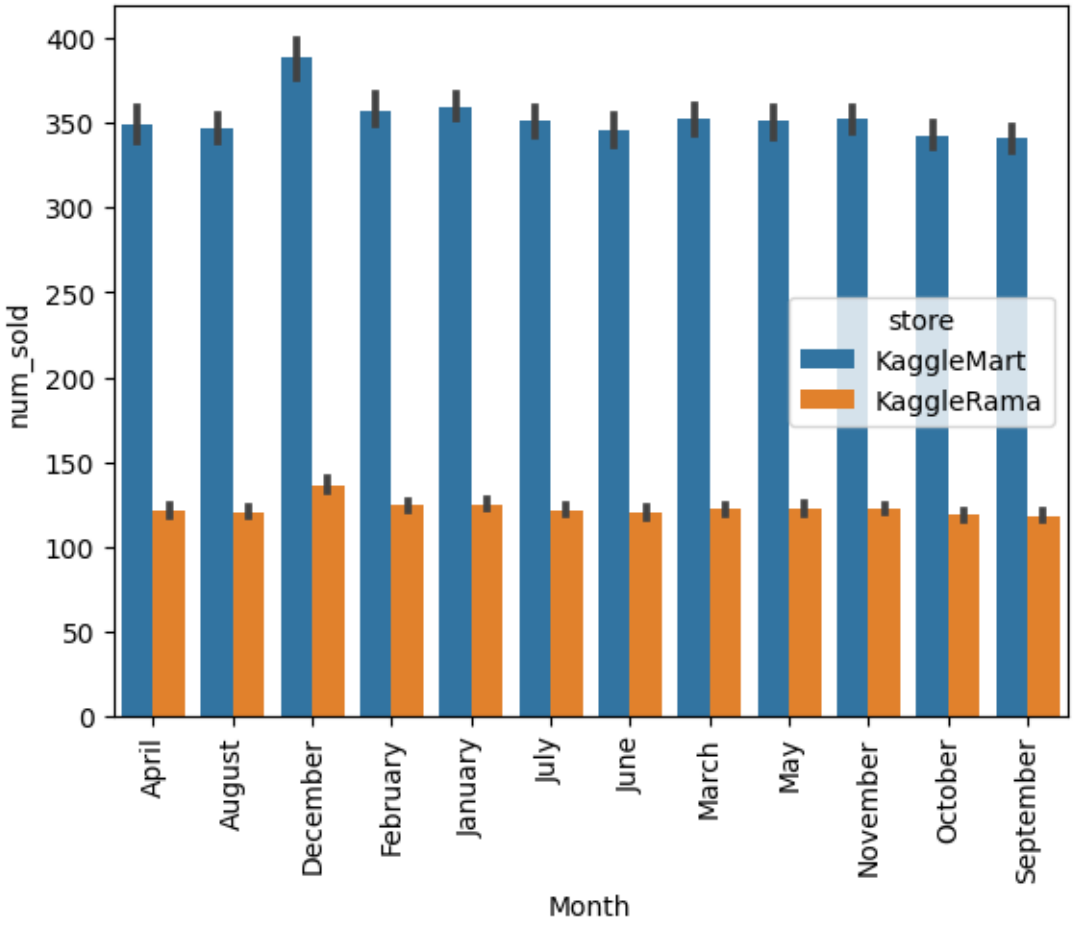


The most number of sales are in the December for Belgium

[49]:
```
Belgium = data[data.country == 'Belgium']
sum5 = Belgium.
 ↪groupby(['Month','store','Day','Date','product','Year'])['num_sold'].sum()
sum5 = pd.DataFrame(sum5)
sum5.reset_index(inplace=True)
sns.barplot(x='Month', y='num_sold', data=sum5, color='blue',␣
 ↪hue='store',palette = ['tab:blue', 'tab:orange'])
plt.xticks(rotation=90)
```

[49]:
```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11]),
 [Text(0, 0, 'April'),
  Text(1, 0, 'August'),
  Text(2, 0, 'December'),
  Text(3, 0, 'February'),
  Text(4, 0, 'January'),
  Text(5, 0, 'July'),
```

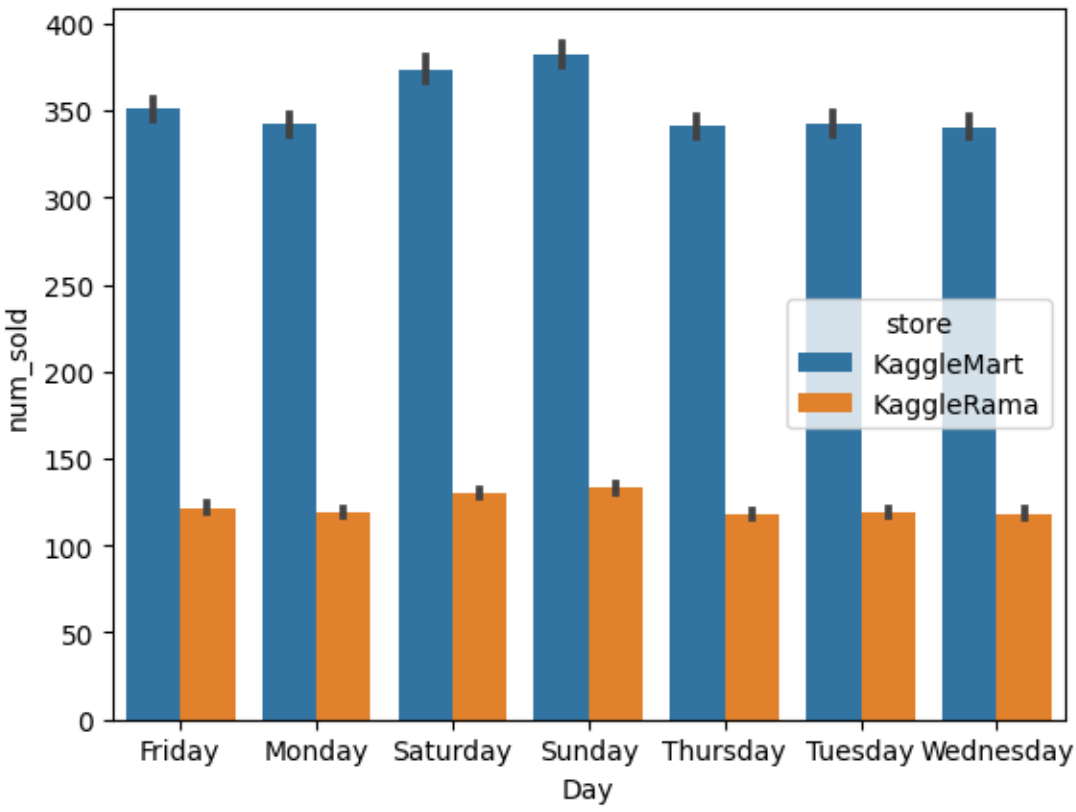
```
Text(6, 0, 'June'),
Text(7, 0, 'March'),
Text(8, 0, 'May'),
Text(9, 0, 'November'),
Text(10, 0, 'October'),
Text(11, 0, 'September')])
```



Most of the sales are done on Sunday and Saturday in the store of Kaggle Mart in Belgium

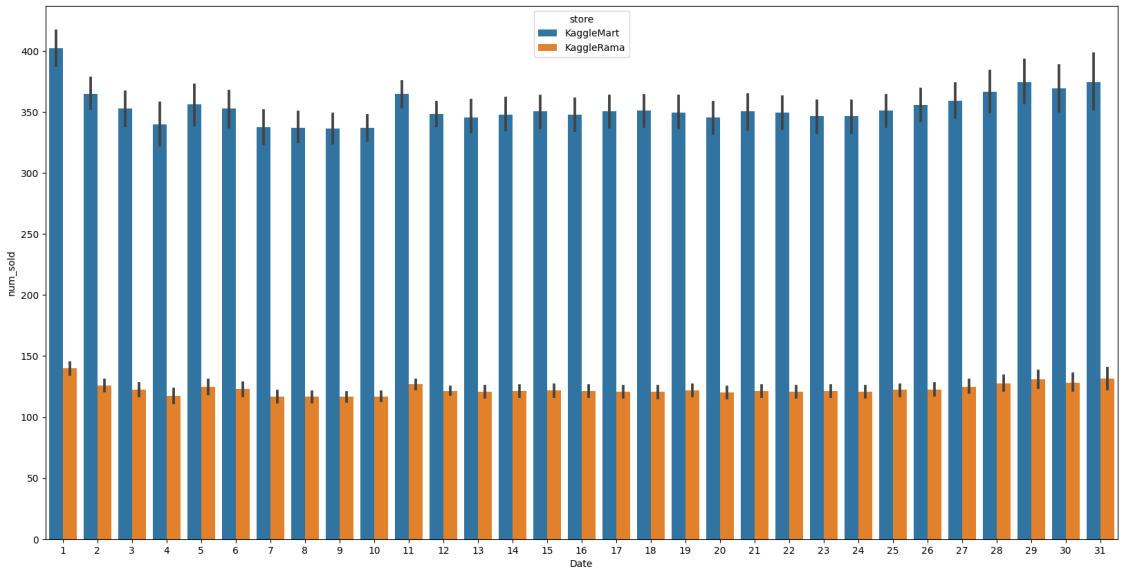
[26]:
```
sns.barplot(x='Day', y='num_sold', data=sum5, color='blue', hue='store',palette␣
↪= ['tab:blue', 'tab:orange'])
```

[26]:
```
<AxesSubplot: xlabel='Day', ylabel='num_sold'>
```



[29]:
```
plt.figure(figsize=(20,10))
sns.barplot(x='Date', y='num_sold', data=sum5, color='blue',␣
↪hue='store',palette = ['tab:blue', 'tab:orange'])
```

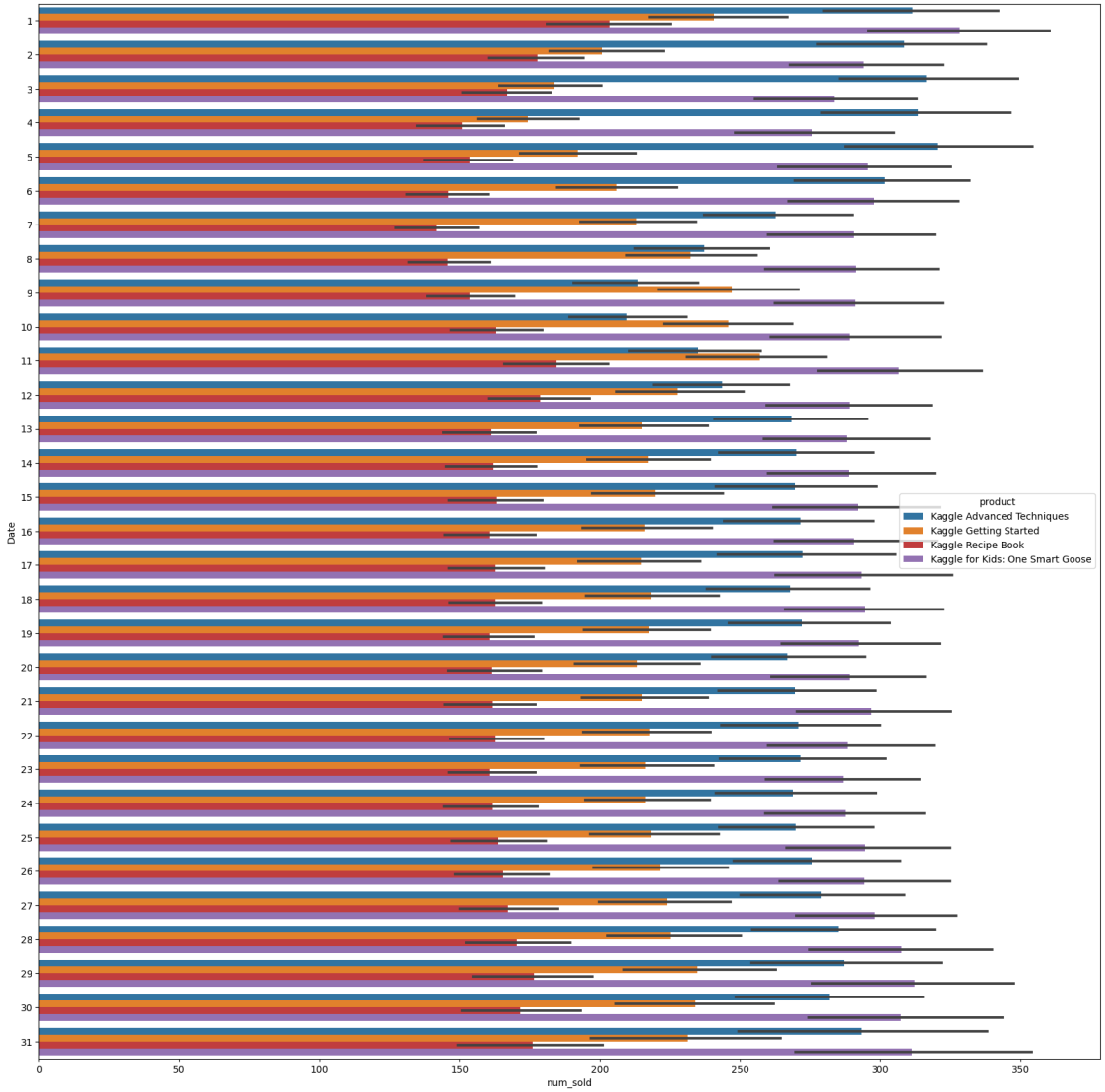
[29]: <AxesSubplot: xlabel='Date', ylabel='num_sold'>



Most of the shopping is averagely done on the first day of the month in belgium

[39]:
```
plt.figure(figsize=(20,20))
sns.barplot(x='num_sold', y='Date', data=sum5, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'],␣
↪orient= 'h')
```

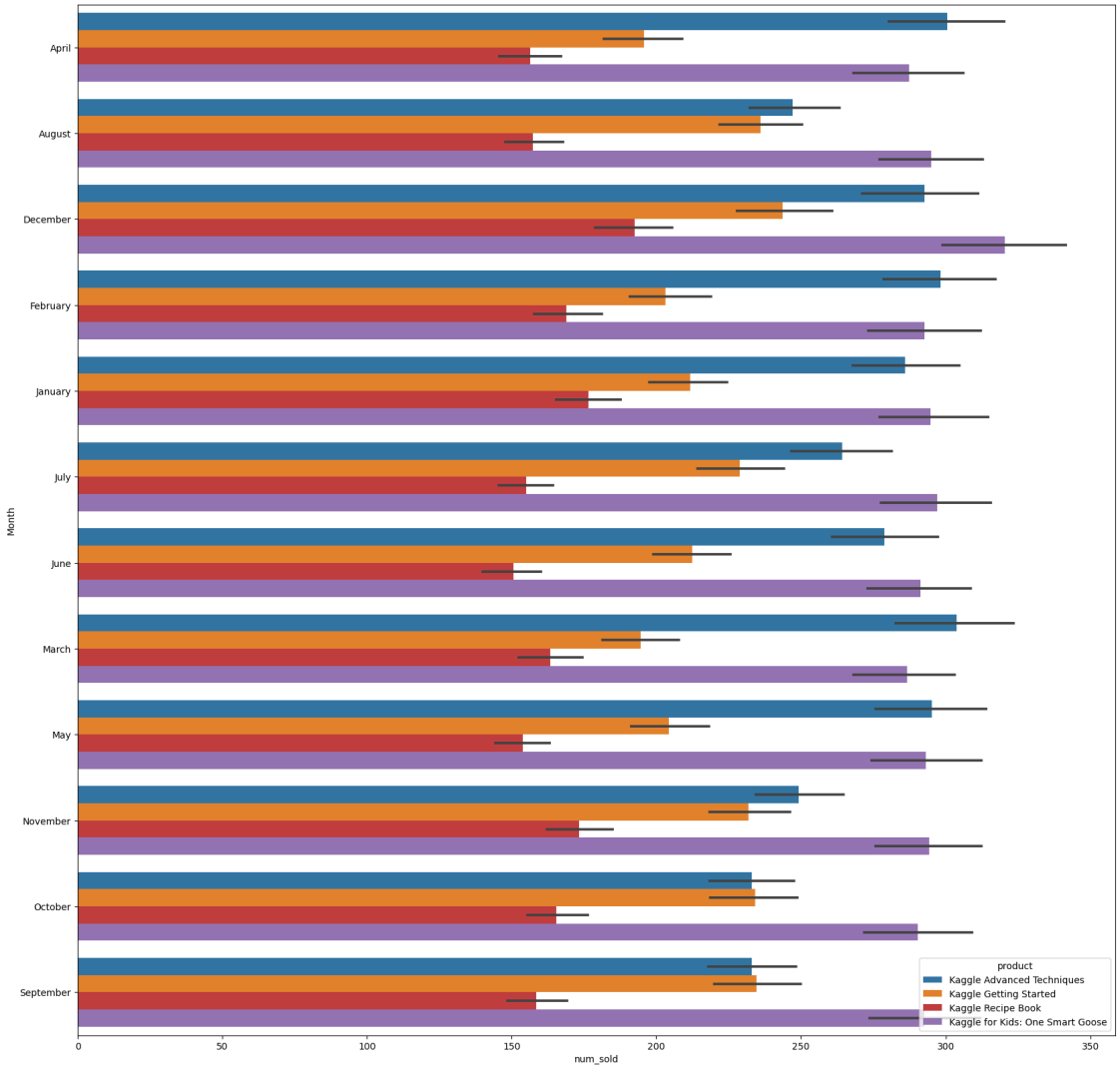
[39]:
```
<AxesSubplot: xlabel='num_sold', ylabel='Date'>
```



Most products are sold on the first day of the month and the product mostly sold are One Smart Goose

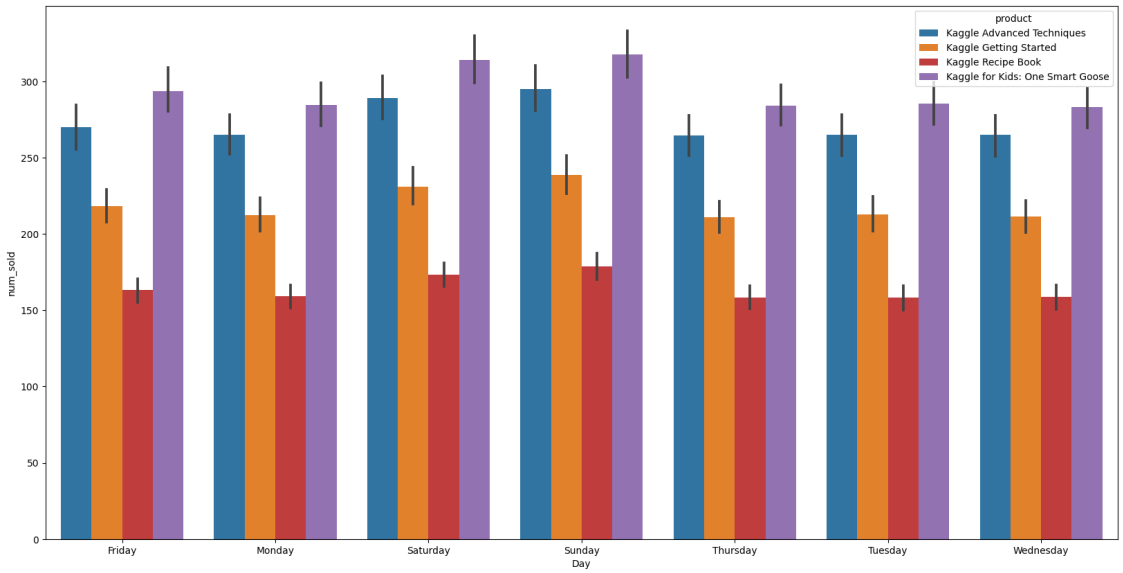
[40]:
```
plt.figure(figsize=(20,20))
sns.barplot(x='num_sold', y='Month', data=sum5, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'],␣
↪orient= 'h')
```

[40]:
```
<AxesSubplot: xlabel='num_sold', ylabel='Month'>
```



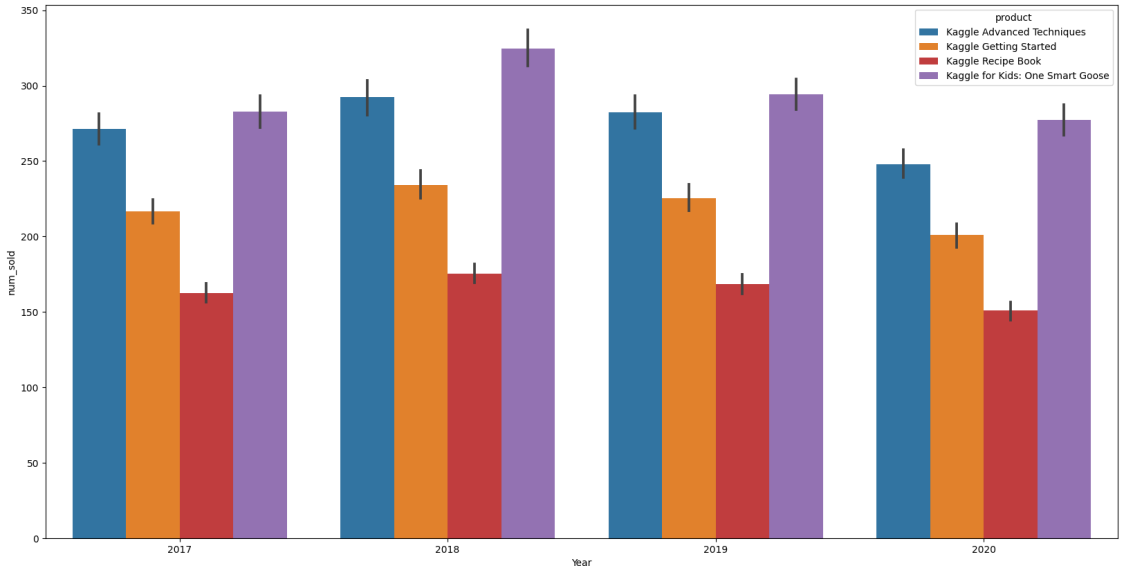
[46]:
```
plt.figure(figsize=(20,10))
sns.barplot(y='num_sold', x='Day', data=sum5, color='blue',␣
 ↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'])
```

[46]:
```
<AxesSubplot: xlabel='Day', ylabel='num_sold'>
```



[52]:
```
plt.figure(figsize=(20,10))
sns.barplot(y='num_sold', x='Year', data=sum5, color='blue',␣
 ↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'])
```

[52]: <AxesSubplot: xlabel='Year', ylabel='num_sold'>

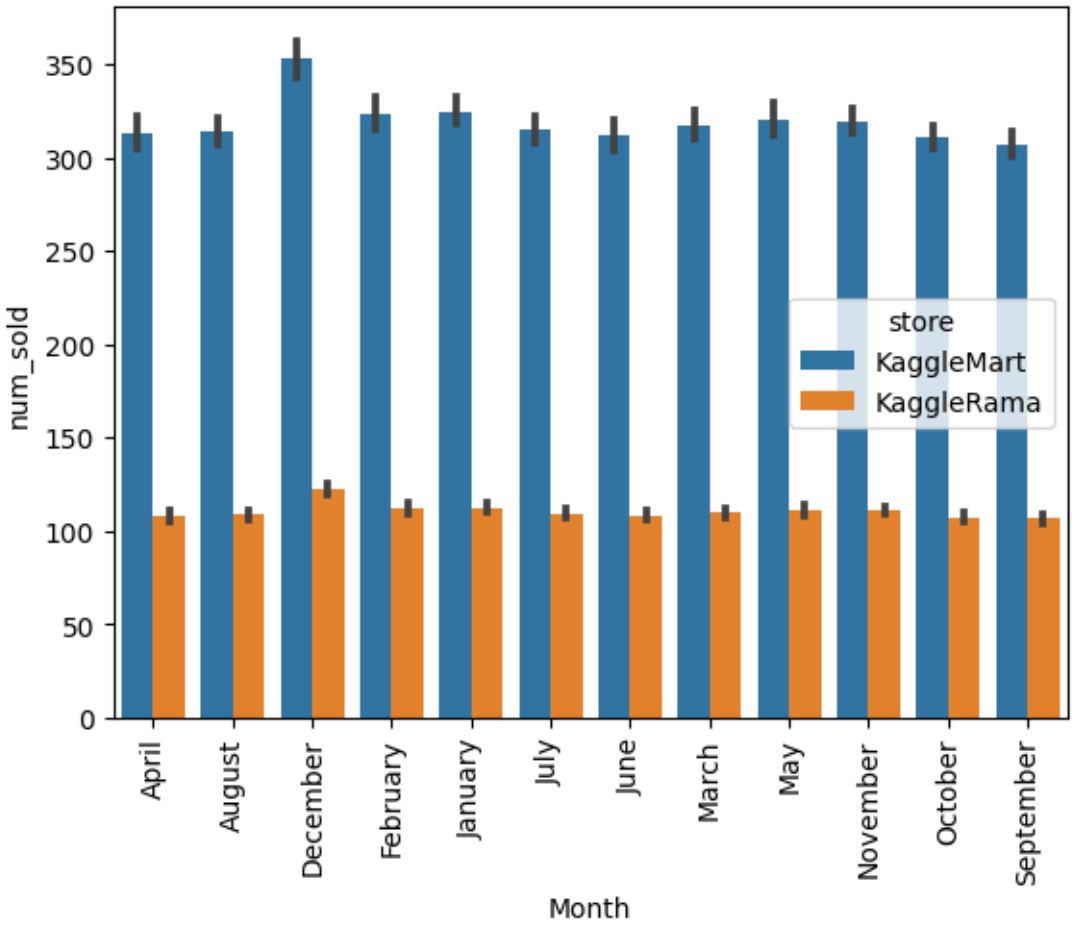


Most of the products were sold in 2018

0.2 France

[53]:
```
Frace = data[data.country == 'France']
sum6 = Frace.
 ↪groupby(['Month','store','Day','product','Date','Year'])['num_sold'].sum()
sum6 = pd.DataFrame(sum6)
sum6.reset_index(inplace=True)
sns.barplot(x='Month', y='num_sold', data=sum6, color='blue',␣
 ↪hue='store',palette = ['tab:blue', 'tab:orange'])
plt.xticks(rotation=90)
```

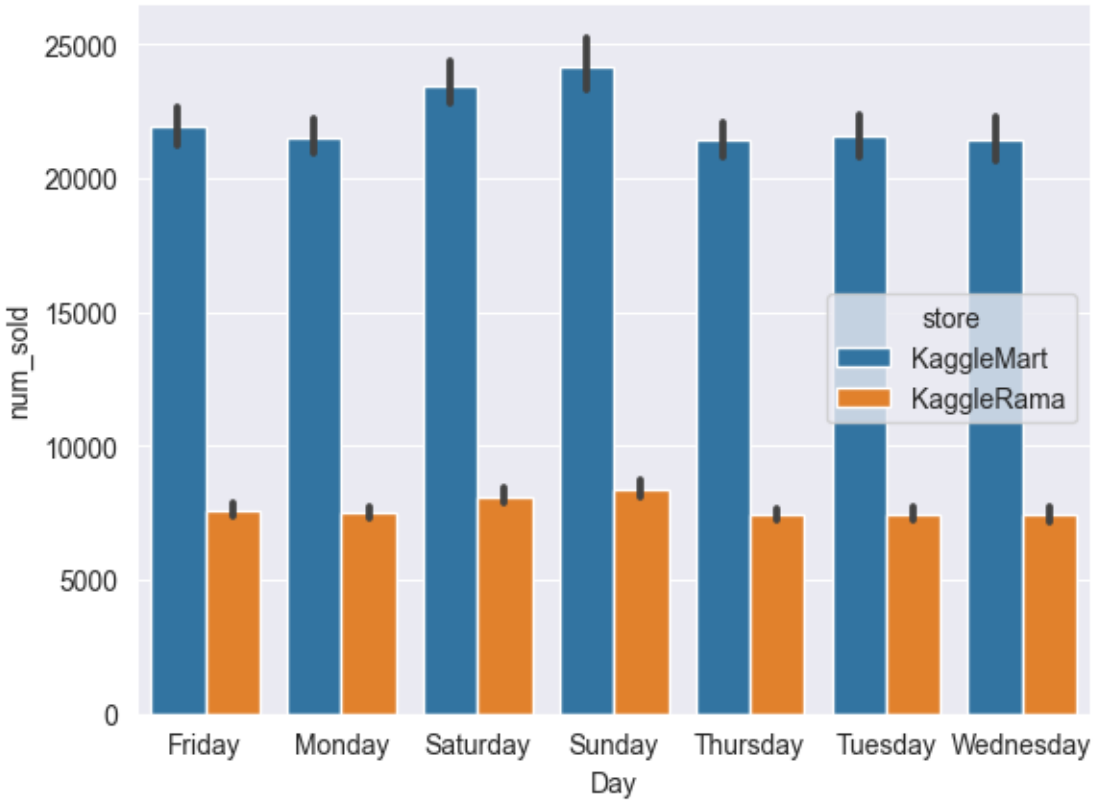
[53]:
```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11]),
 [Text(0, 0, 'April'),
  Text(1, 0, 'August'),
  Text(2, 0, 'December'),
  Text(3, 0, 'February'),
  Text(4, 0, 'January'),
  Text(5, 0, 'July'),
  Text(6, 0, 'June'),
  Text(7, 0, 'March'),
  Text(8, 0, 'May'),
  Text(9, 0, 'November'),
  Text(10, 0, 'October'),
  Text(11, 0, 'September')])
```



The most number of products are sold in December in France that too from the Kaggle Mart store

[21]:
```
sns.barplot(x='Day', y='num_sold', data=sum6, color='blue', hue='store',palette␣
↪= ['tab:blue', 'tab:orange'])
```

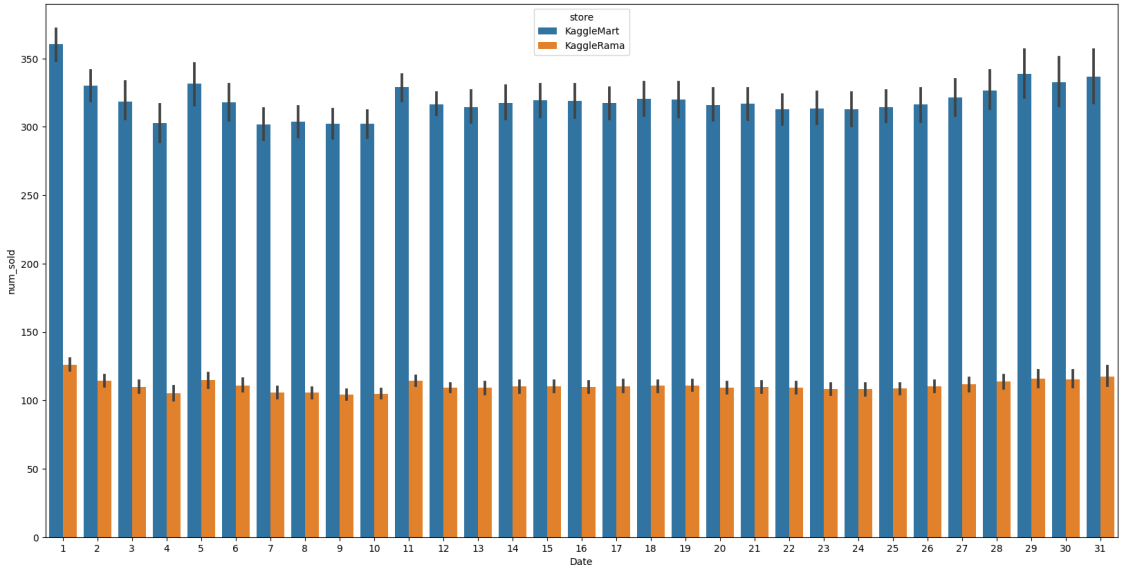
[21]:
```
<AxesSubplot: xlabel='Day', ylabel='num_sold'>
```



The most number of products are sold on Saturday and Sunday in France

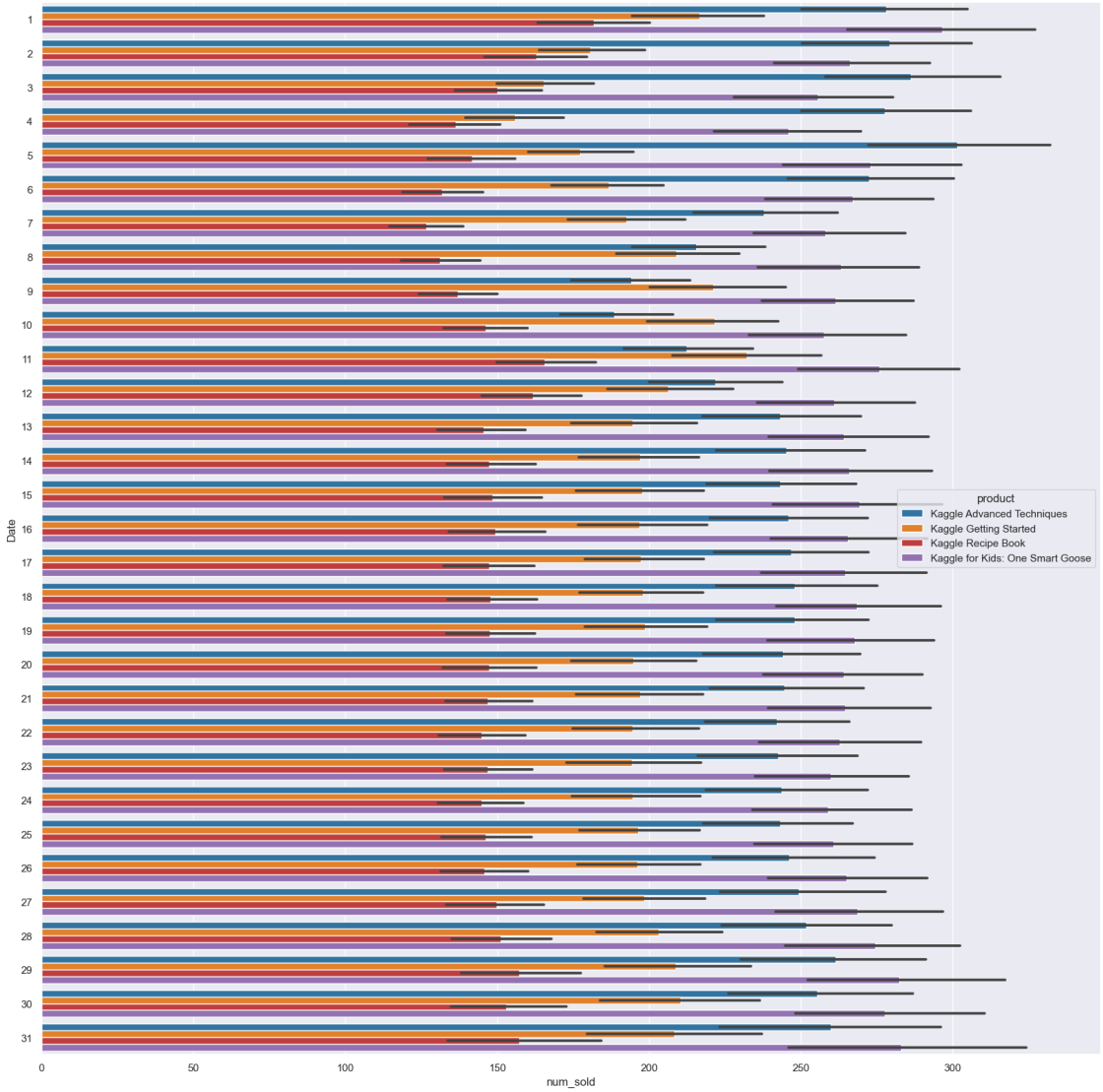
[54]:
```
plt.figure(figsize=(20,10))
sns.barplot(x='Date', y='num_sold', data=sum6, color='blue',␣
 ↪hue='store',palette = ['tab:blue', 'tab:orange'])
```

[54]:
```
<AxesSubplot: xlabel='Date', ylabel='num_sold'>
```



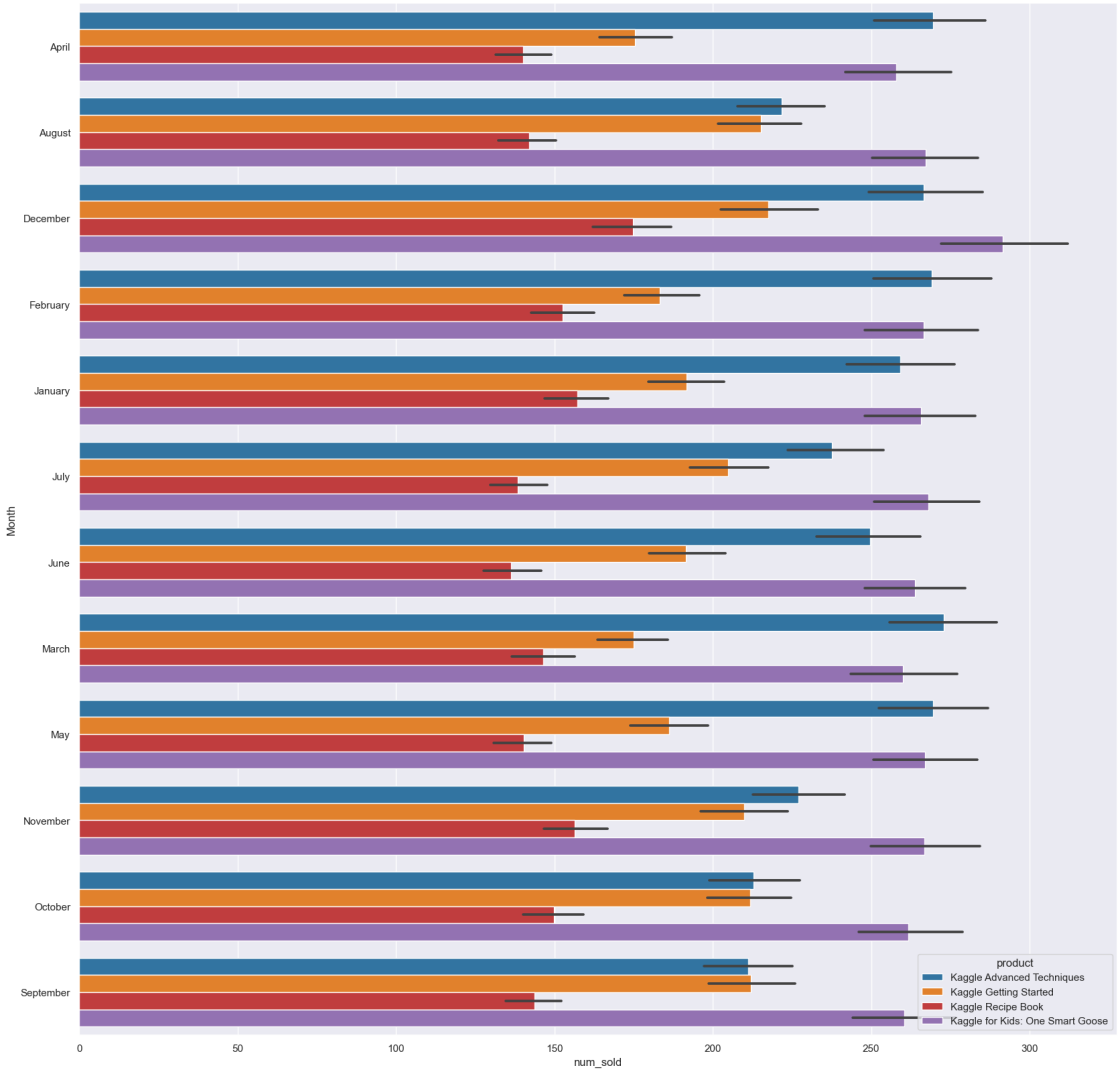
[59]:
```
plt.figure(figsize=(20,20))
sns.set_theme(style="darkgrid")
sns.barplot(x='num_sold', y='Date', data=sum6, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'],␣
↪orient= 'h')
```

[59]:
```
<AxesSubplot: xlabel='num_sold', ylabel='Date'>
```



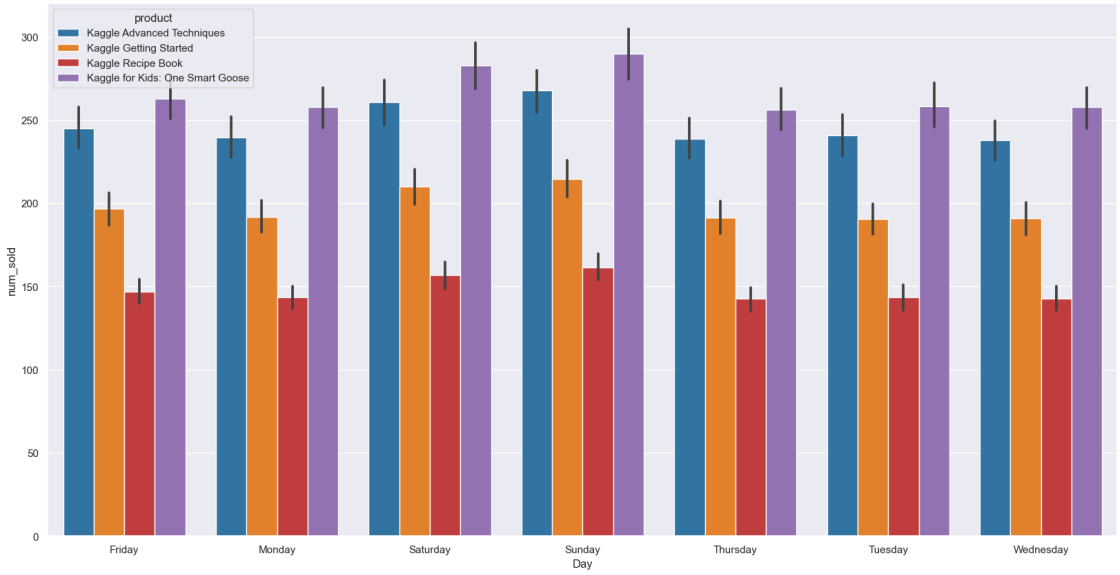
[58]:
```
plt.figure(figsize=(20,20))
sns.barplot(x='num_sold', y='Month', data=sum6, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'],␣
↪orient= 'h')
```

[58]: <AxesSubplot: xlabel='num_sold', ylabel='Month'>



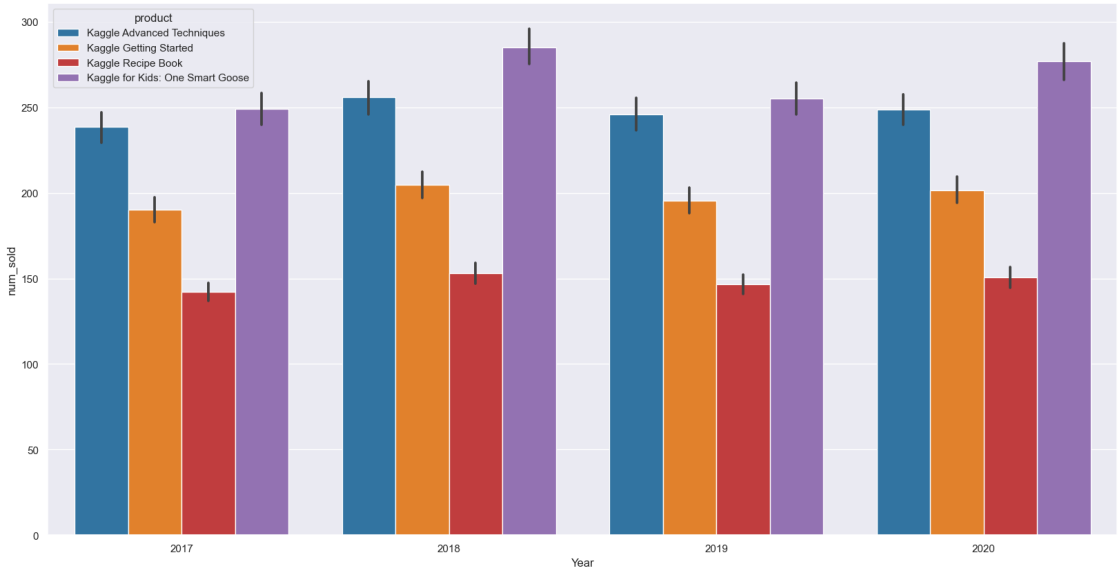
[60]:
```
plt.figure(figsize=(20,10))
sns.barplot(y='num_sold', x='Day', data=sum6, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'])
```

[60]:
```
<AxesSubplot: xlabel='Day', ylabel='num_sold'>
```



[61]:
```
plt.figure(figsize=(20,10))
sns.barplot(y='num_sold', x='Year', data=sum6, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'])
```

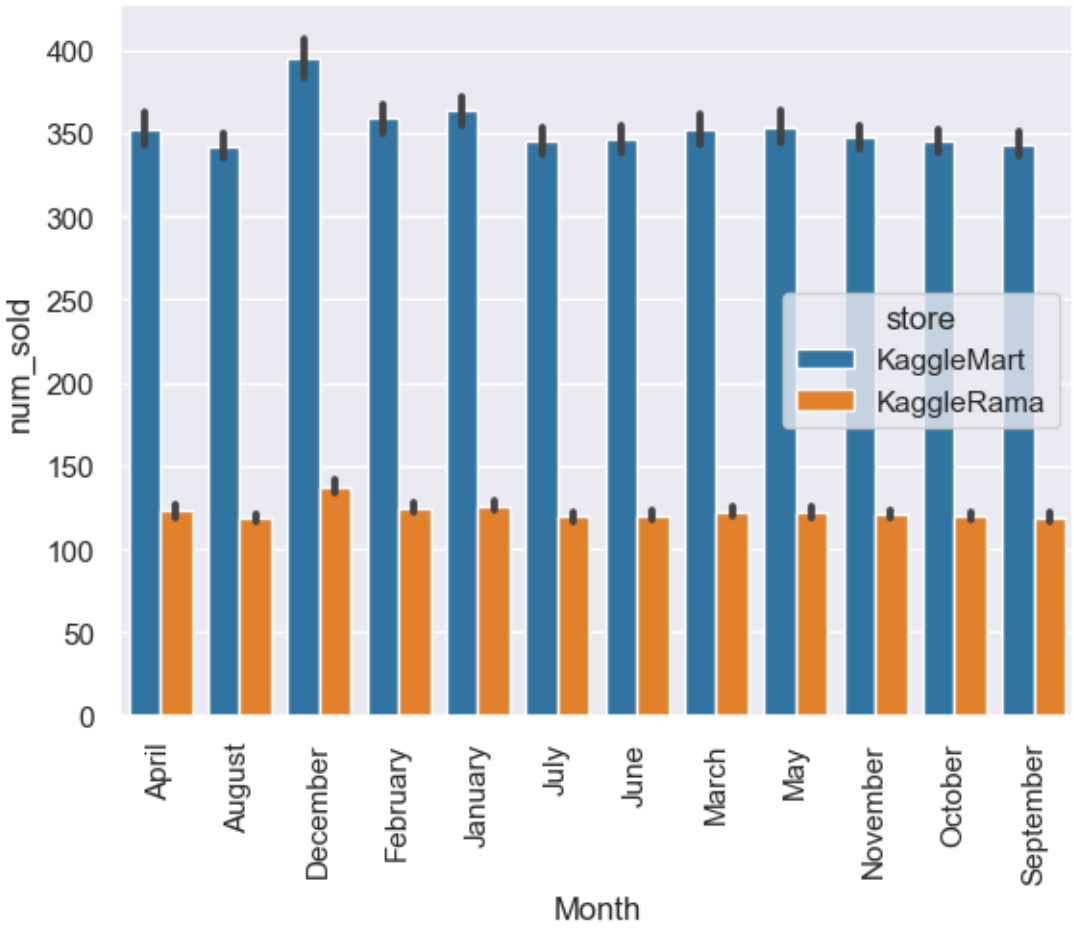
[61]: <AxesSubplot: xlabel='Year', ylabel='num_sold'>



0.3 Germany

[62]:
```
Germany = data[data.country == 'Germany']
sum7 = Germany.
 ↪groupby(['Month','store','Day','Year','Date','product'])['num_sold'].sum()
sum7 = pd.DataFrame(sum7)
sum7.reset_index(inplace=True)
sns.barplot(x='Month', y='num_sold', data=sum7, color='blue',␣
 ↪hue='store',palette = ['tab:blue', 'tab:orange'])
plt.xticks(rotation=90)
```

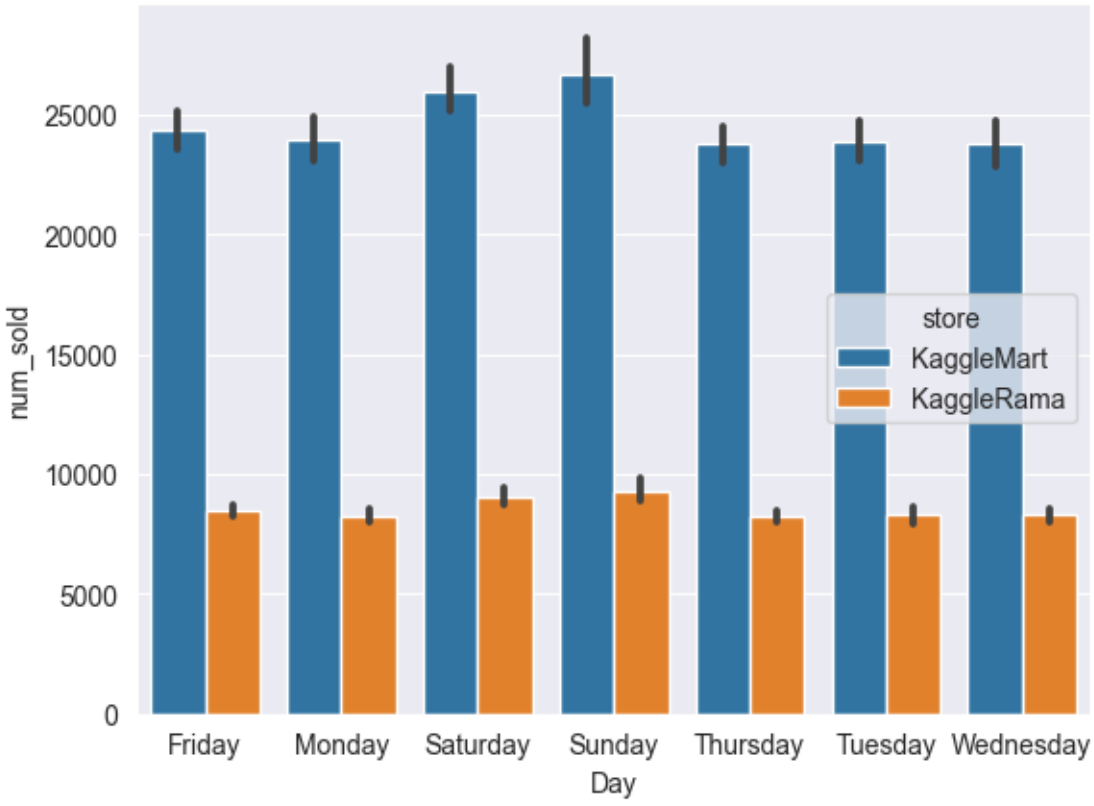
[62]:
```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11]),
 [Text(0, 0, 'April'),
  Text(1, 0, 'August'),
  Text(2, 0, 'December'),
  Text(3, 0, 'February'),
  Text(4, 0, 'January'),
  Text(5, 0, 'July'),
  Text(6, 0, 'June'),
  Text(7, 0, 'March'),
  Text(8, 0, 'May'),
  Text(9, 0, 'November'),
  Text(10, 0, 'October'),
  Text(11, 0, 'September')])
```



Even in Germany, most number of sales are done in the month of December from Kaagle Mart store

[23]:
```
sns.barplot(x='Day', y='num_sold', data=sum7, color='blue', hue='store',palette␣
↪= ['tab:blue', 'tab:orange'])
```

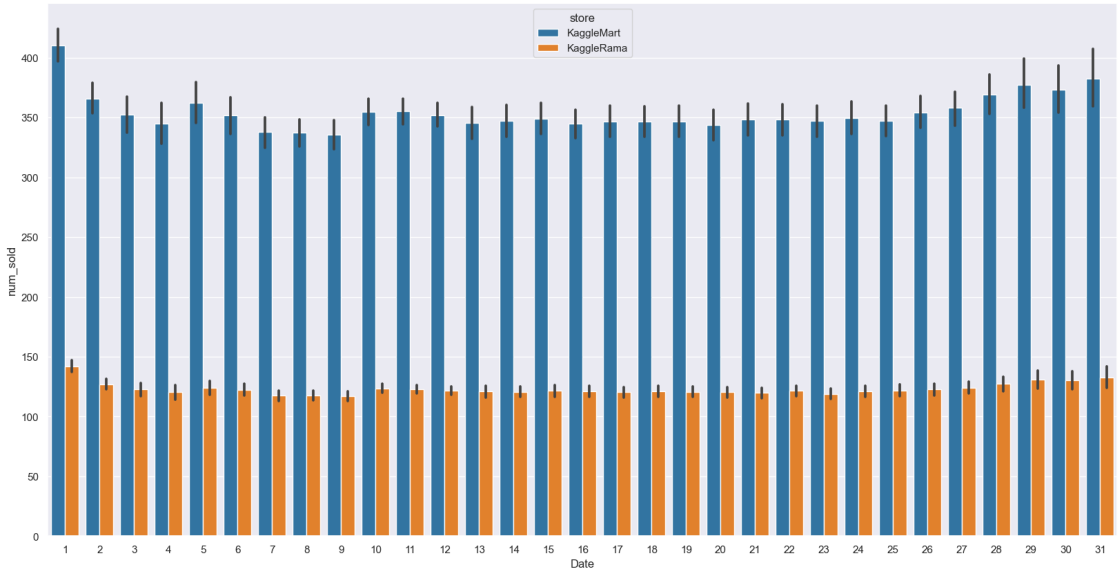
[23]:
```
<AxesSubplot: xlabel='Day', ylabel='num_sold'>
```



The most number of products in France are sold from the store of Kaggle mart on Saturday and Sunday

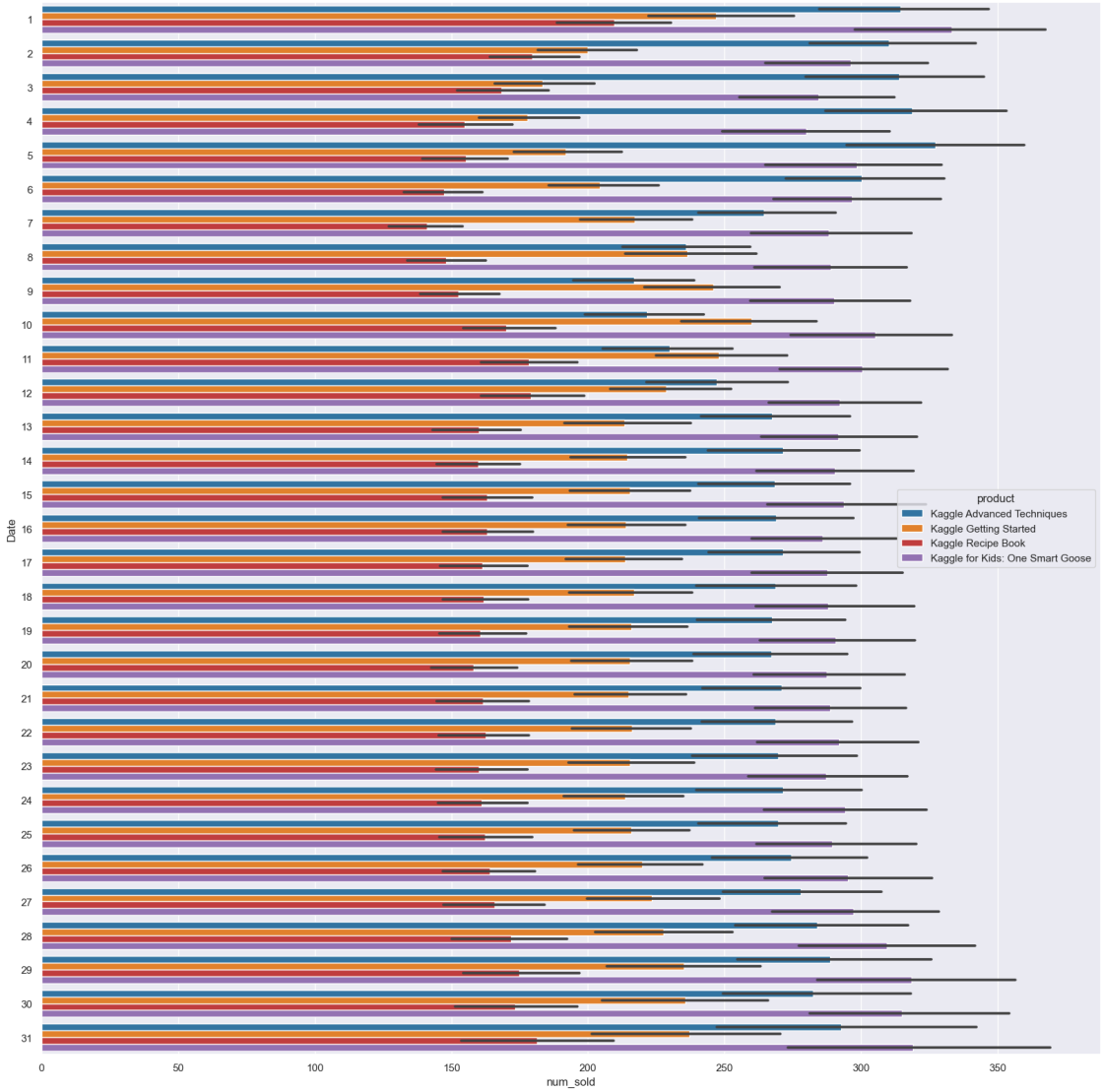
[63]:
```
plt.figure(figsize=(20,10))
sns.barplot(x='Date', y='num_sold', data=sum7, color='blue',␣
 ↪hue='store',palette = ['tab:blue', 'tab:orange'])
```

[63]:
```
<AxesSubplot: xlabel='Date', ylabel='num_sold'>
```



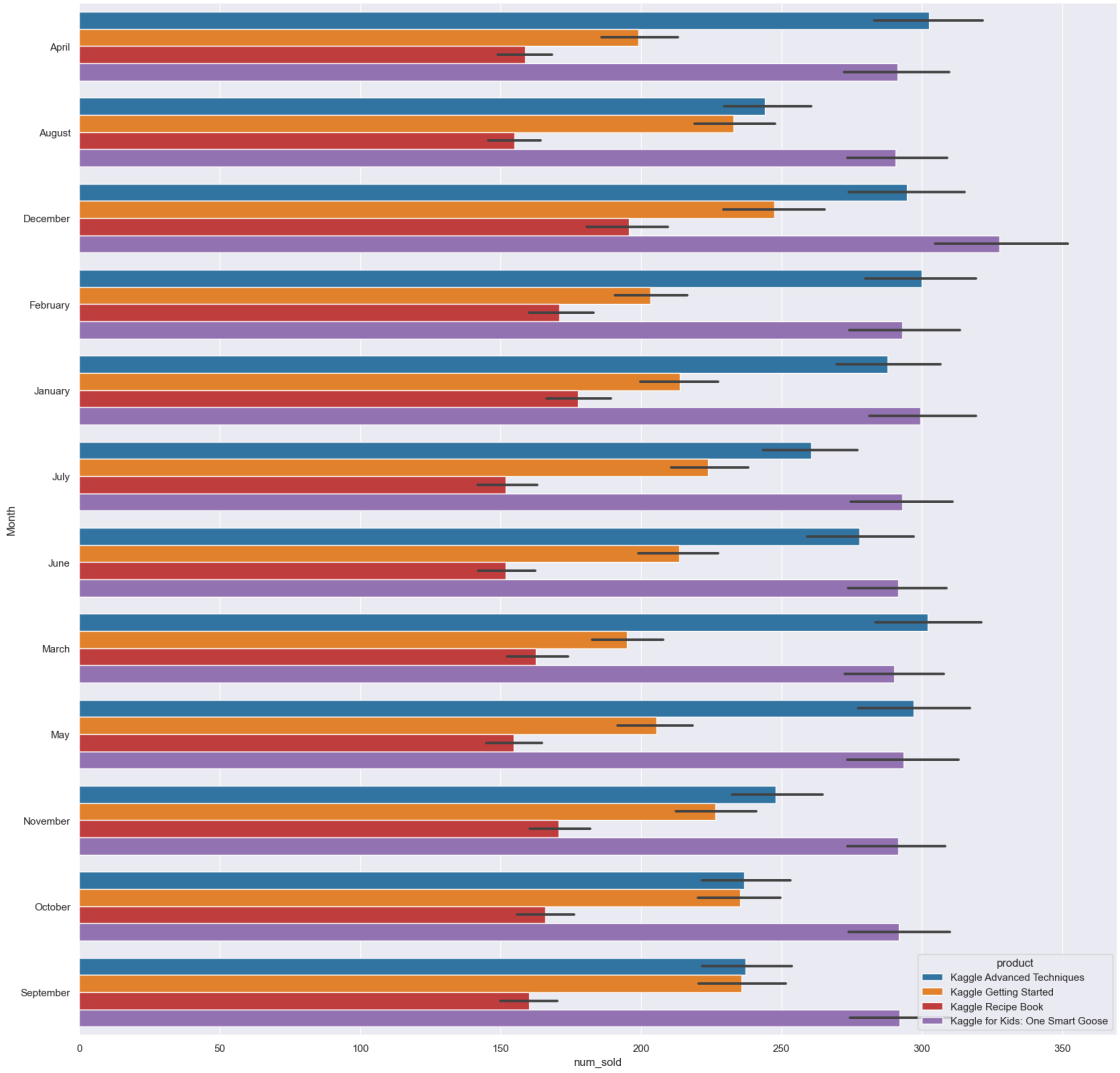
[64]:
```
plt.figure(figsize=(20,20))
sns.set_theme(style="darkgrid")
sns.barplot(x='num_sold', y='Date', data=sum7, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'],␣
↪orient= 'h')
```

[64]:
```
<AxesSubplot: xlabel='num_sold', ylabel='Date'>
```



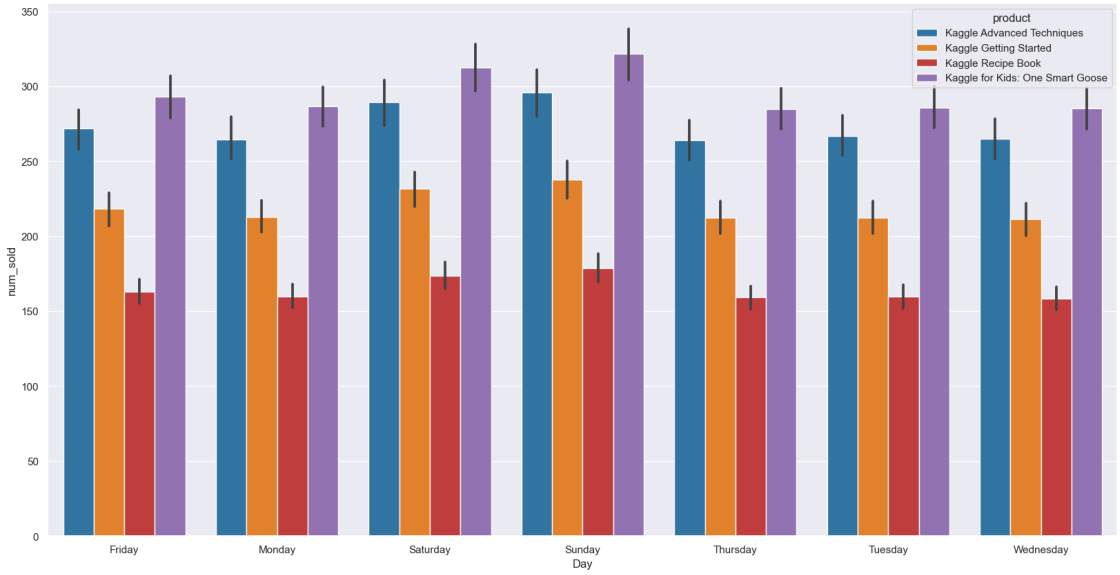
[65]:
```
plt.figure(figsize=(20,20))
sns.barplot(x='num_sold', y='Month', data=sum7, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'],␣
↪orient= 'h')
```

[65]: <AxesSubplot: xlabel='num_sold', ylabel='Month'>



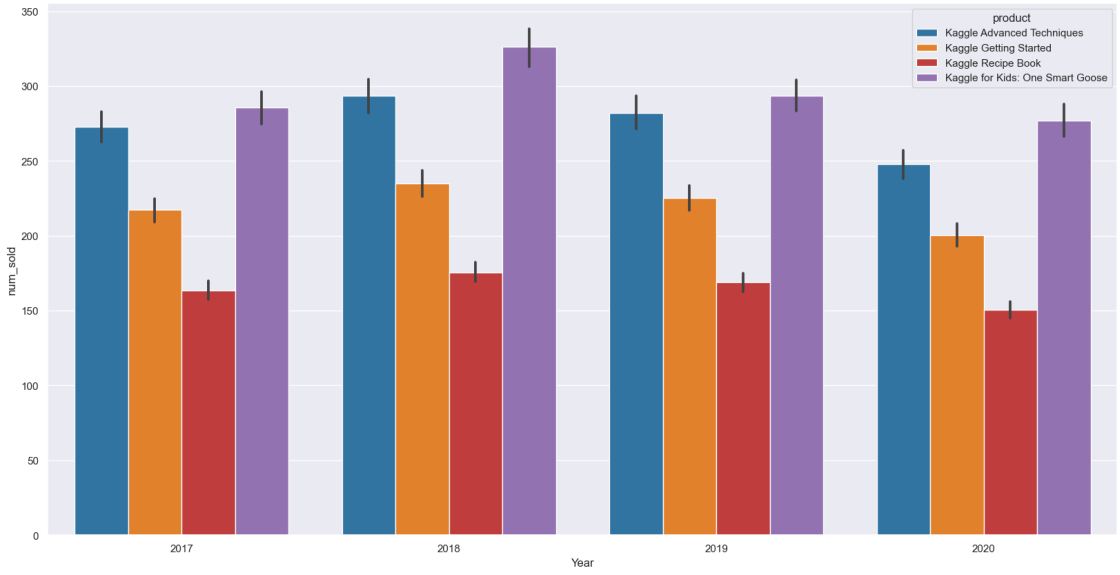
[66]:
```
plt.figure(figsize=(20,10))
sns.barplot(y='num_sold', x='Day', data=sum7, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'])
```

[66]:
```
<AxesSubplot: xlabel='Day', ylabel='num_sold'>
```



[67]:
```
plt.figure(figsize=(20,10))
sns.barplot(y='num_sold', x='Year', data=sum7, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'])
```

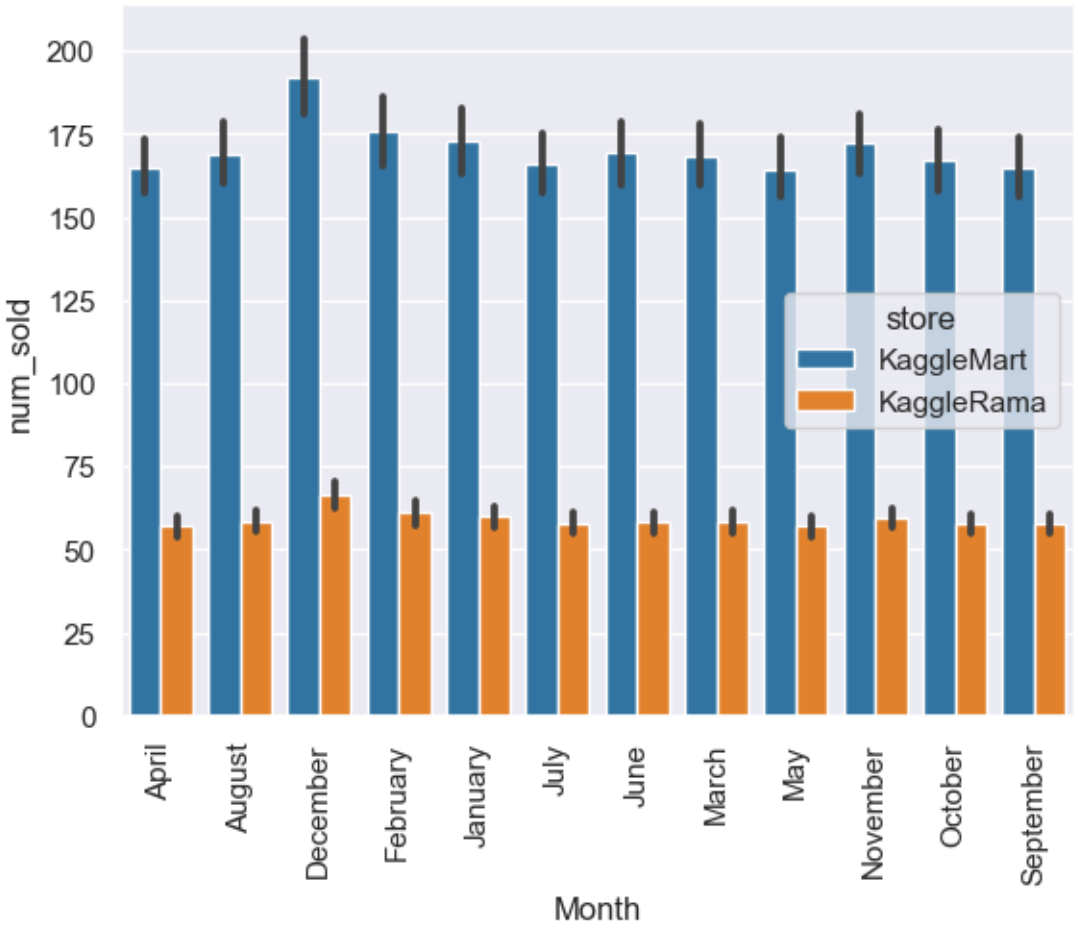
[67]: <AxesSubplot: xlabel='Year', ylabel='num_sold'>



0.4 Poland

[69]:
```
Poland = data[data.country == 'Poland']
sum8 = Poland.
 ↪groupby(['Month','store','Day','Year','product','Date'])['num_sold'].sum()
sum8 = pd.DataFrame(sum8)
sum8.reset_index(inplace=True)
sns.barplot(x='Month', y='num_sold', data=sum8, color='blue',␣
 ↪hue='store',palette = ['tab:blue', 'tab:orange'])
plt.xticks(rotation=90)
```

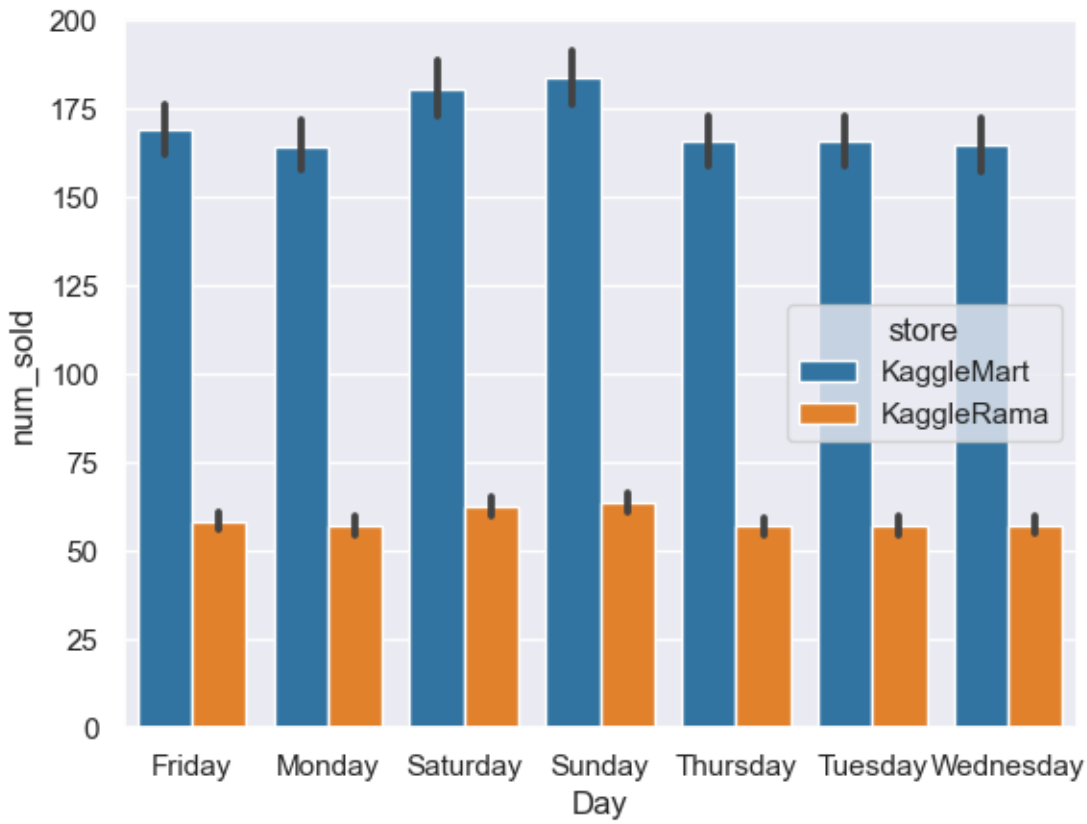
[69]:
```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11]),
 [Text(0, 0, 'April'),
  Text(1, 0, 'August'),
  Text(2, 0, 'December'),
  Text(3, 0, 'February'),
  Text(4, 0, 'January'),
  Text(5, 0, 'July'),
  Text(6, 0, 'June'),
  Text(7, 0, 'March'),
  Text(8, 0, 'May'),
  Text(9, 0, 'November'),
  Text(10, 0, 'October'),
  Text(11, 0, 'September')])
```



The most number of products in Poland are sold in December from Kaagle Mart store

[70]:
```
sns.barplot(x='Day', y='num_sold', data=sum8, color='blue', hue='store',palette␣
↪= ['tab:blue', 'tab:orange'])
```

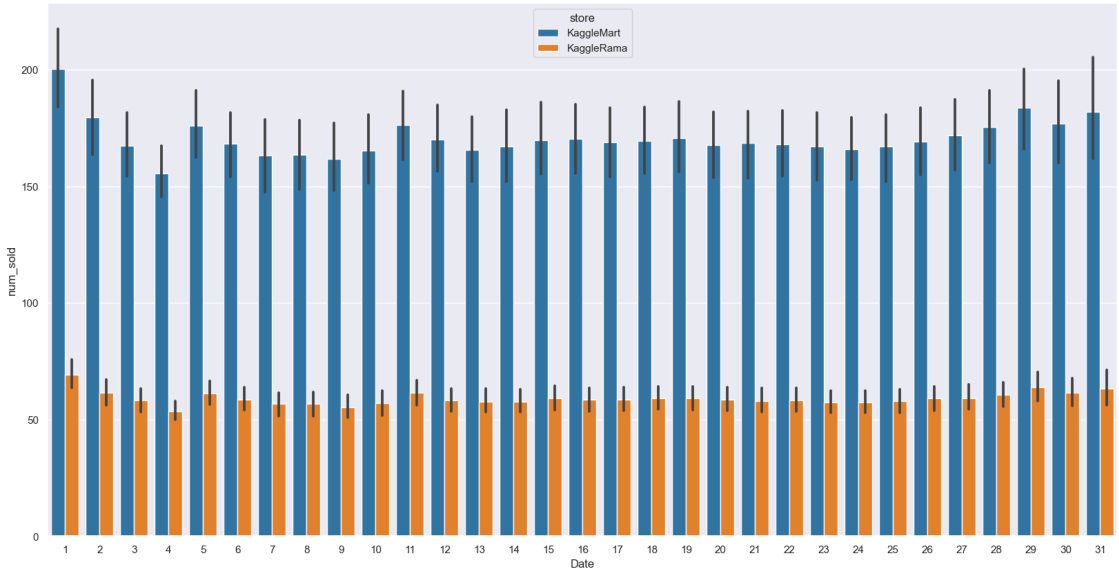
[70]:
```
<AxesSubplot: xlabel='Day', ylabel='num_sold'>
```



Even in Poland, the most number of products are sold on the days of Saturday and Sunady from Kaggle Mart store

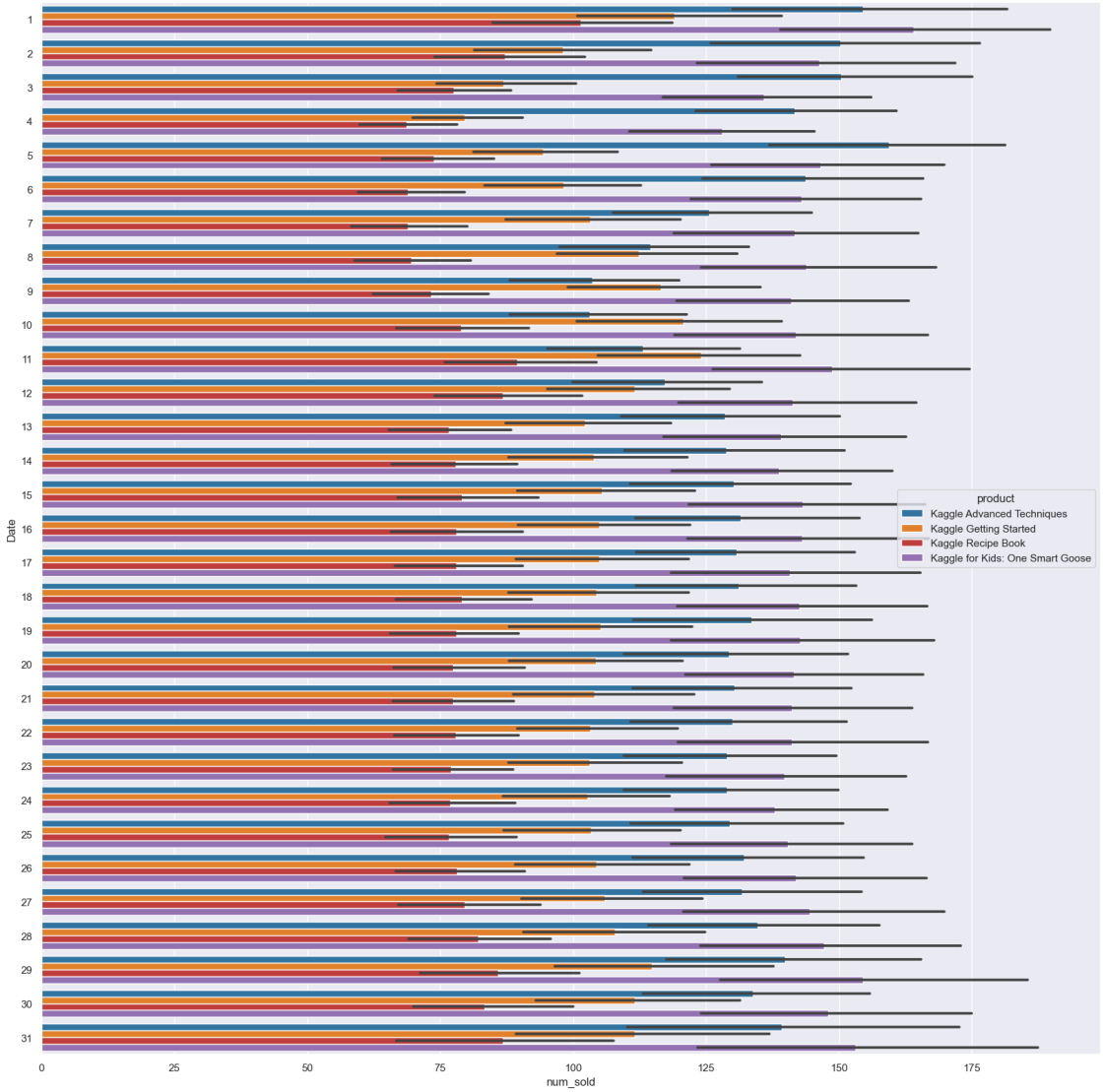
[71]:
```
plt.figure(figsize=(20,10))
sns.barplot(x='Date', y='num_sold', data=sum8, color='blue',␣
↪hue='store',palette = ['tab:blue', 'tab:orange'])
```

[71]:
```
<AxesSubplot: xlabel='Date', ylabel='num_sold'>
```



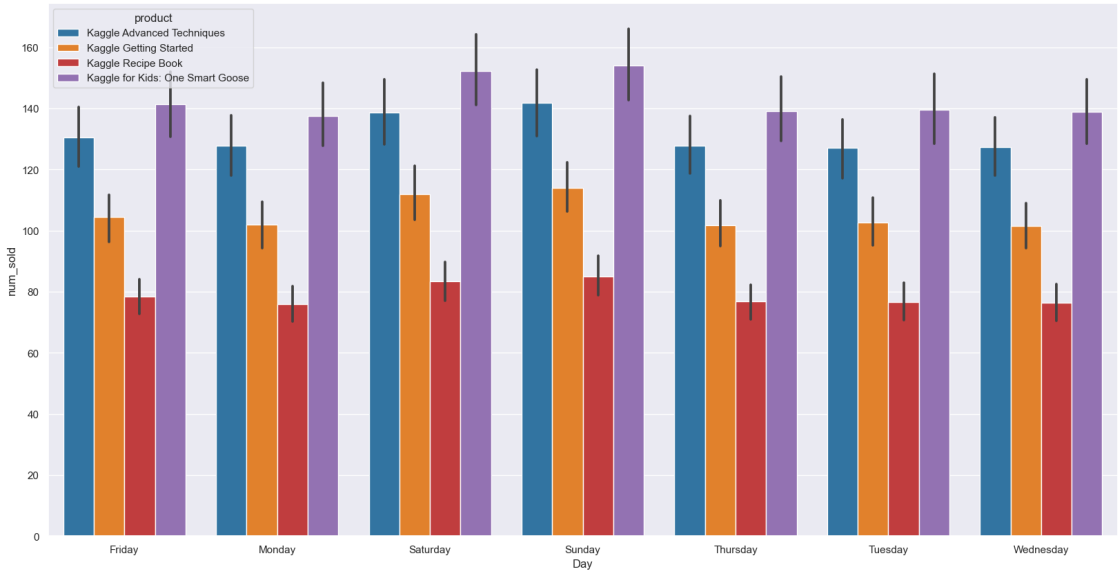
[74]:
```
plt.figure(figsize=(20,20))
sns.set_theme(style="darkgrid")
sns.barplot(x='num_sold', y='Date', data=sum8, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'],␣
↪orient= 'h')
```

[74]: <AxesSubplot: xlabel='num_sold', ylabel='Date'>



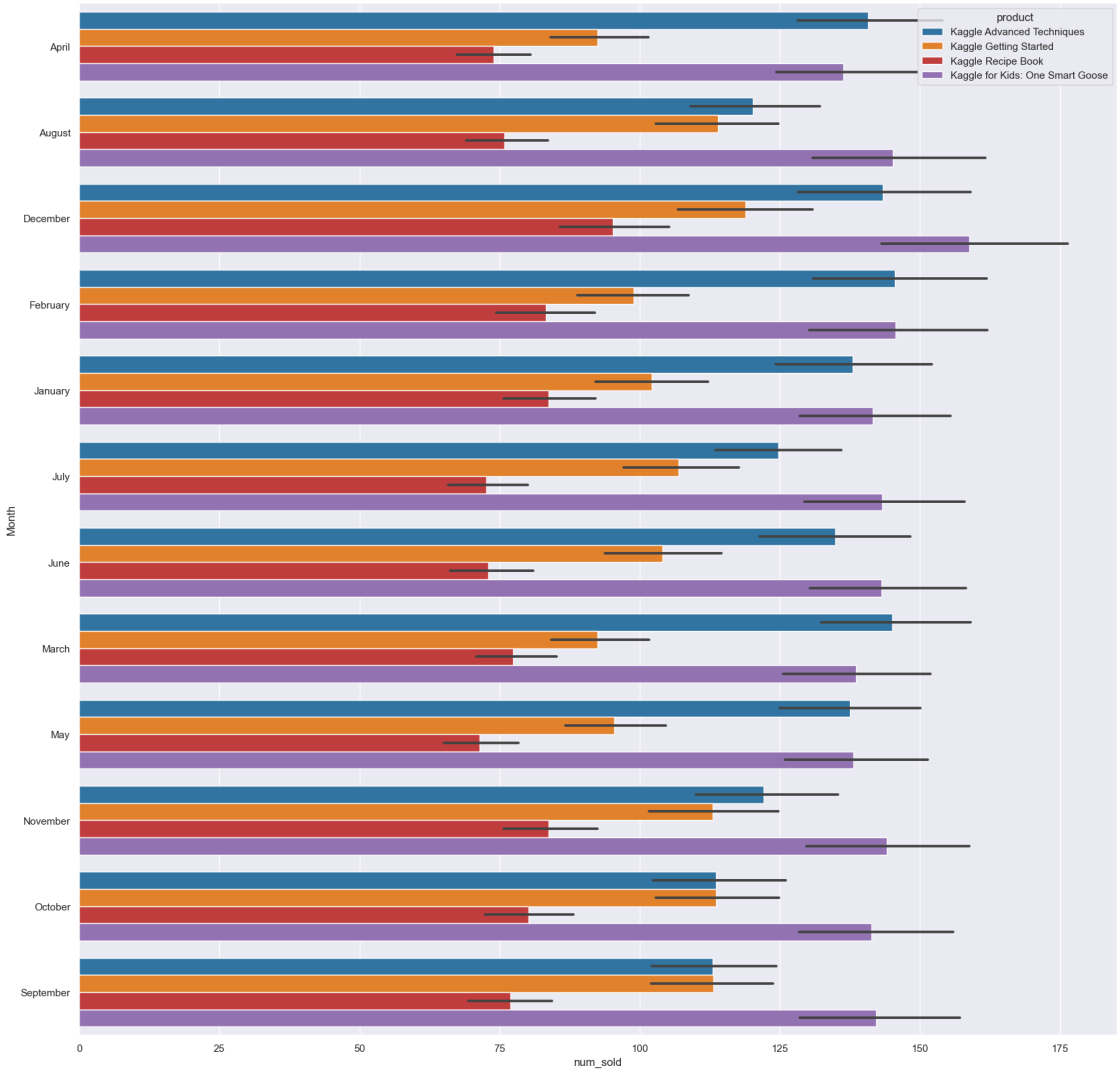
[85]:
```
plt.figure(figsize=(20,10))
sns.barplot(y='num_sold', x='Day', data=sum8, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'])
```

[85]:
```
<AxesSubplot: xlabel='Day', ylabel='num_sold'>
```



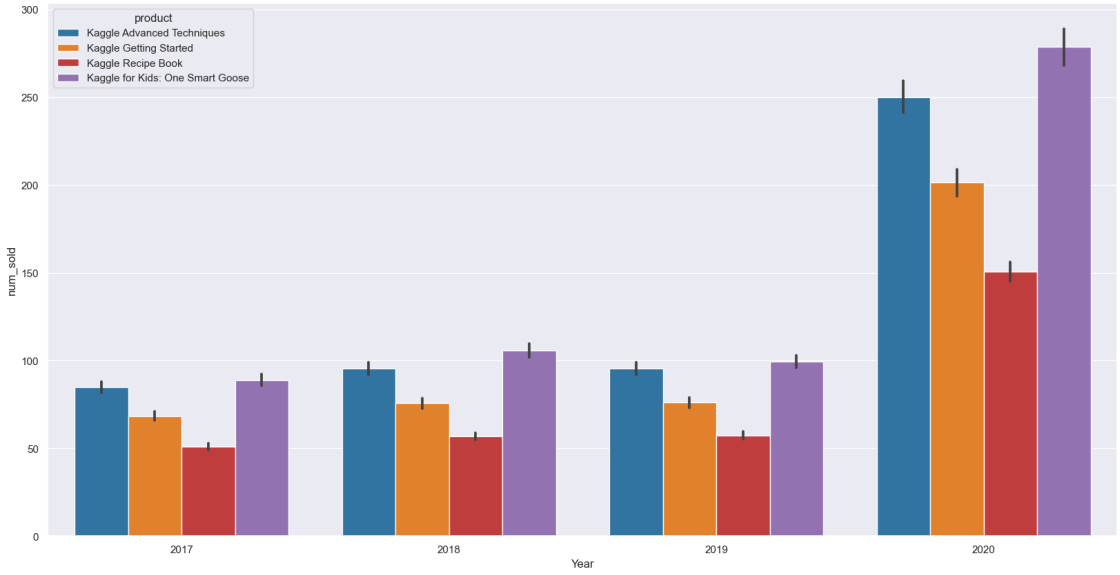
[75]:
```
plt.figure(figsize=(20,20))
sns.barplot(x='num_sold', y='Month', data=sum8, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'],␣
↪orient= 'h')
```

[75]:
```
<AxesSubplot: xlabel='num_sold', ylabel='Month'>
```



[76]:
```
plt.figure(figsize=(20,10))
sns.barplot(y='num_sold', x='Year', data=sum8, color='blue',␣
 ↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'])
```

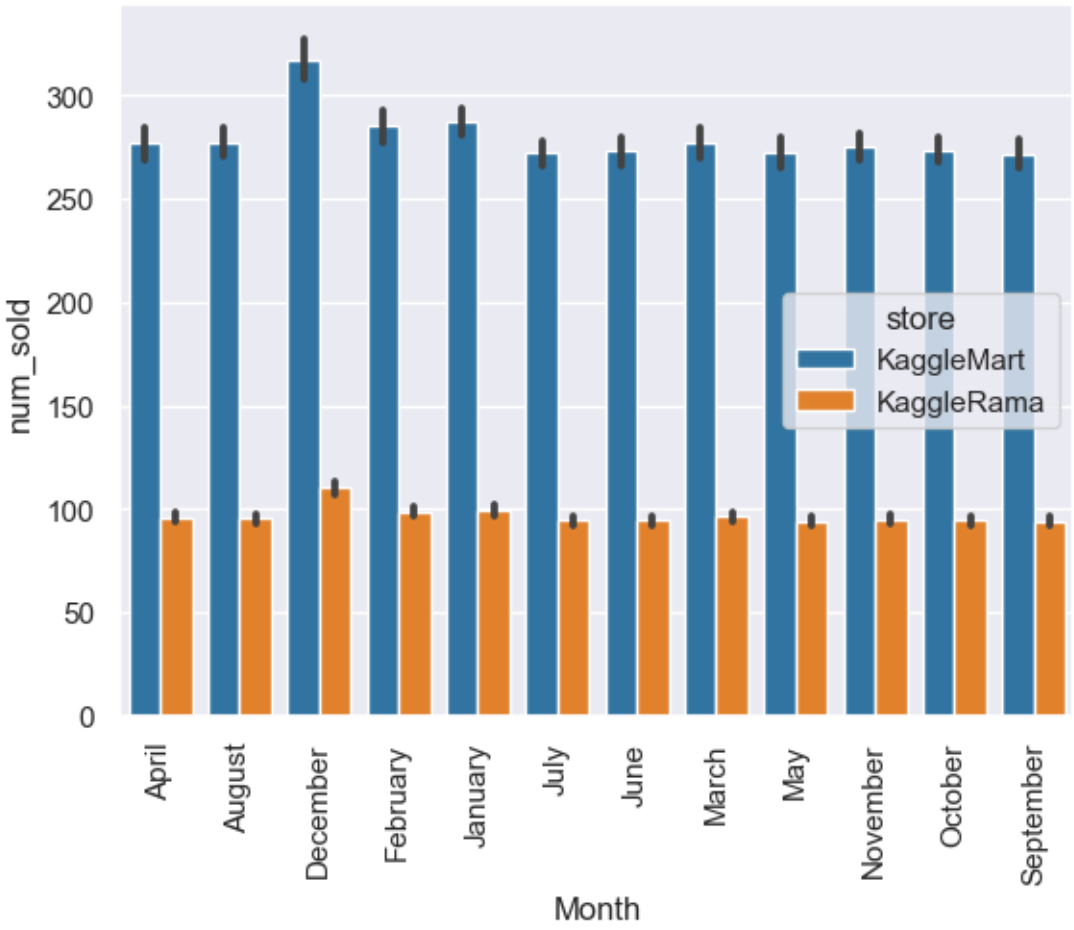
[76]:
```
<AxesSubplot: xlabel='Year', ylabel='num_sold'>
```



0.5 Italy

[86]:
```
Italy = data[data.country == 'Italy']
sum9 = Italy.
 ↪groupby(['Month','store','Day','Year','product','Date'])['num_sold'].sum()
sum9 = pd.DataFrame(sum9)
sum9.reset_index(inplace=True)
sns.barplot(x='Month', y='num_sold', data=sum9, color='blue',␣
 ↪hue='store',palette = ['tab:blue', 'tab:orange'])
plt.xticks(rotation=90)
```

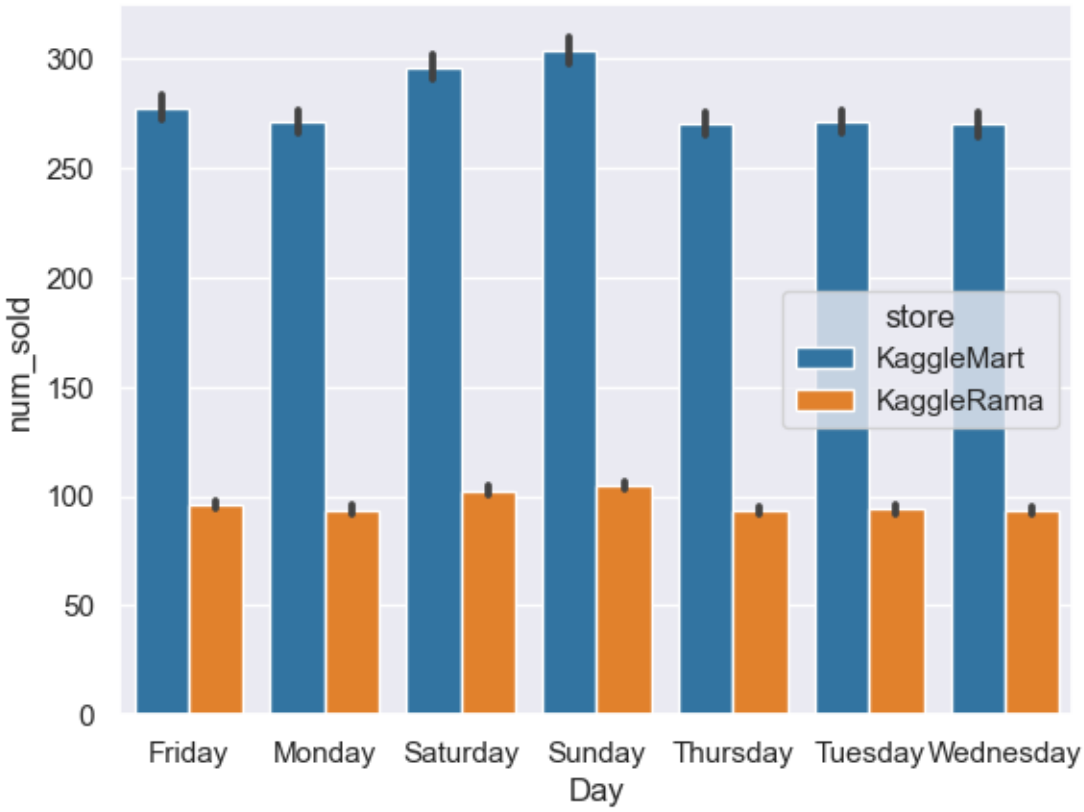
[86]:
```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11]),
 [Text(0, 0, 'April'),
  Text(1, 0, 'August'),
  Text(2, 0, 'December'),
  Text(3, 0, 'February'),
  Text(4, 0, 'January'),
  Text(5, 0, 'July'),
  Text(6, 0, 'June'),
  Text(7, 0, 'March'),
  Text(8, 0, 'May'),
  Text(9, 0, 'November'),
  Text(10, 0, 'October'),
  Text(11, 0, 'September')])
```



The most number of products are sold in the month of December in Italy from the Kaggle Mart store and least in February

[87]:
```
sns.barplot(x='Day', y='num_sold', data=sum9, color='blue', hue='store',palette␣
↪= ['tab:blue', 'tab:orange'])
```

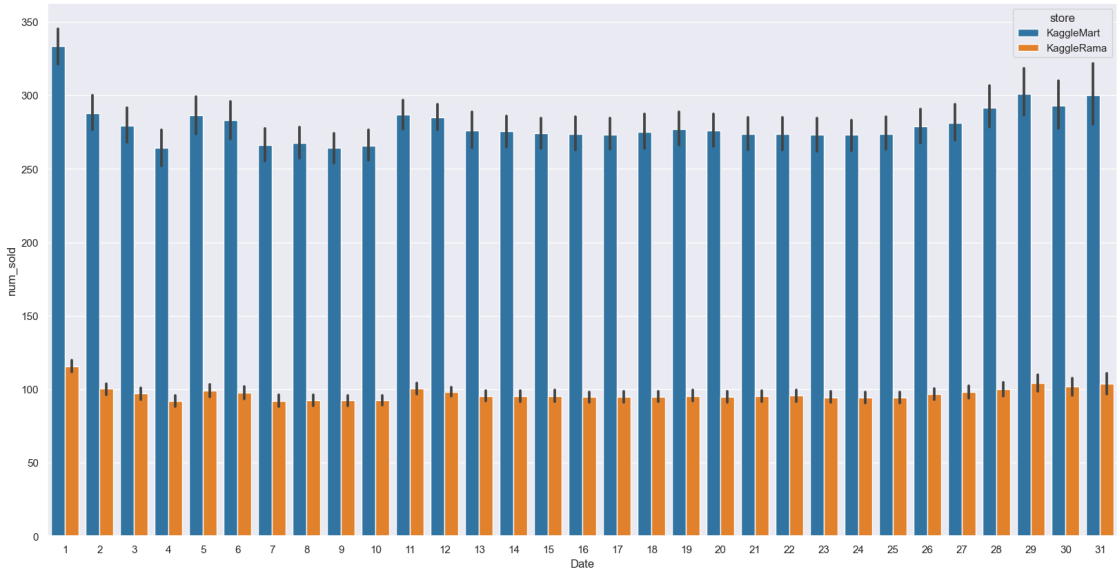
[87]:
```
<AxesSubplot: xlabel='Day', ylabel='num_sold'>
```



The most number of products are sold on Saturday and Sunday even in Italy from Kaggle Mart store

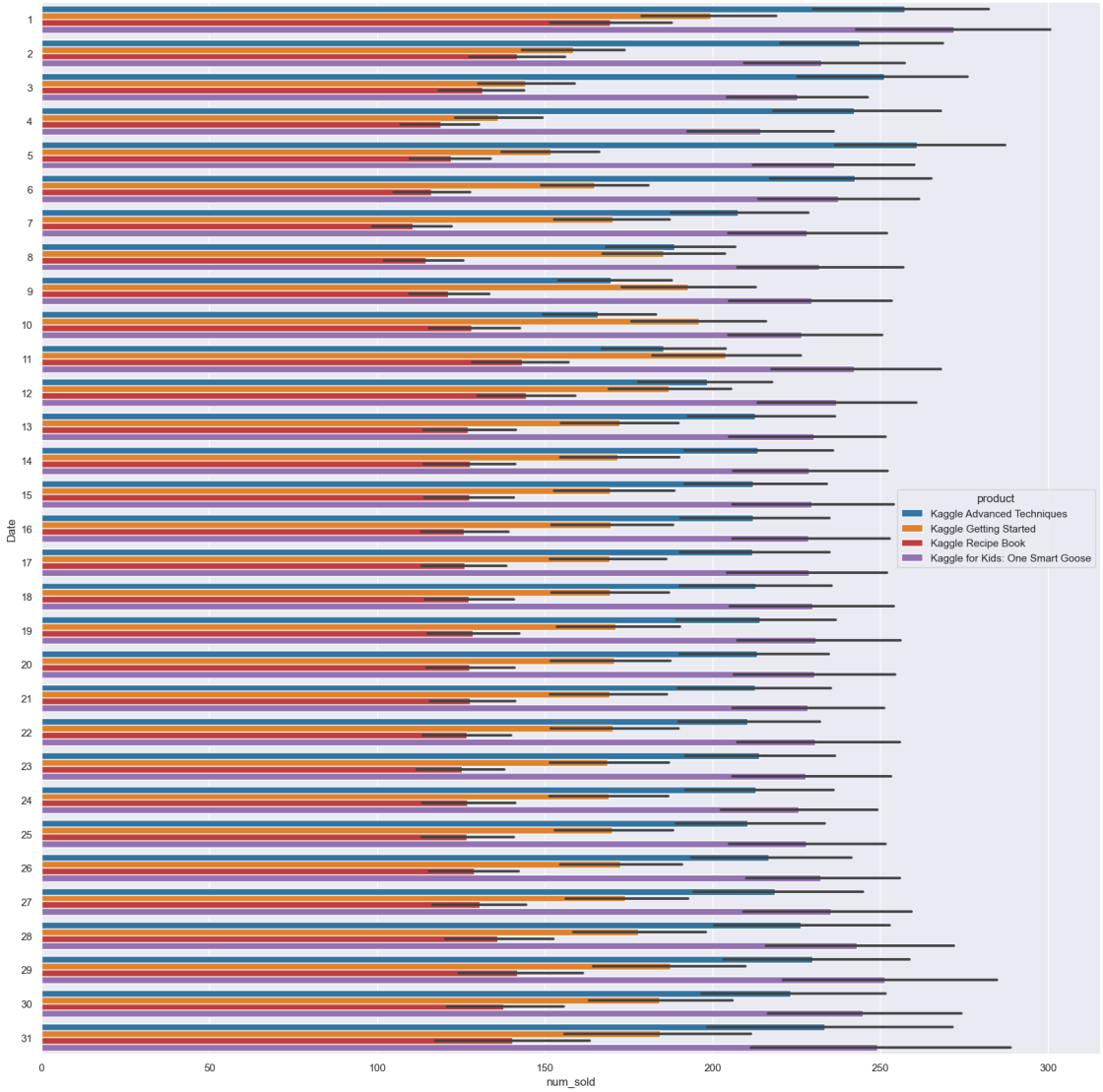
[88]:
```
plt.figure(figsize=(20,10))
sns.barplot(x='Date', y='num_sold', data=sum9, color='blue',␣
 ↪hue='store',palette = ['tab:blue', 'tab:orange'])
```

[88]:
```
<AxesSubplot: xlabel='Date', ylabel='num_sold'>
```



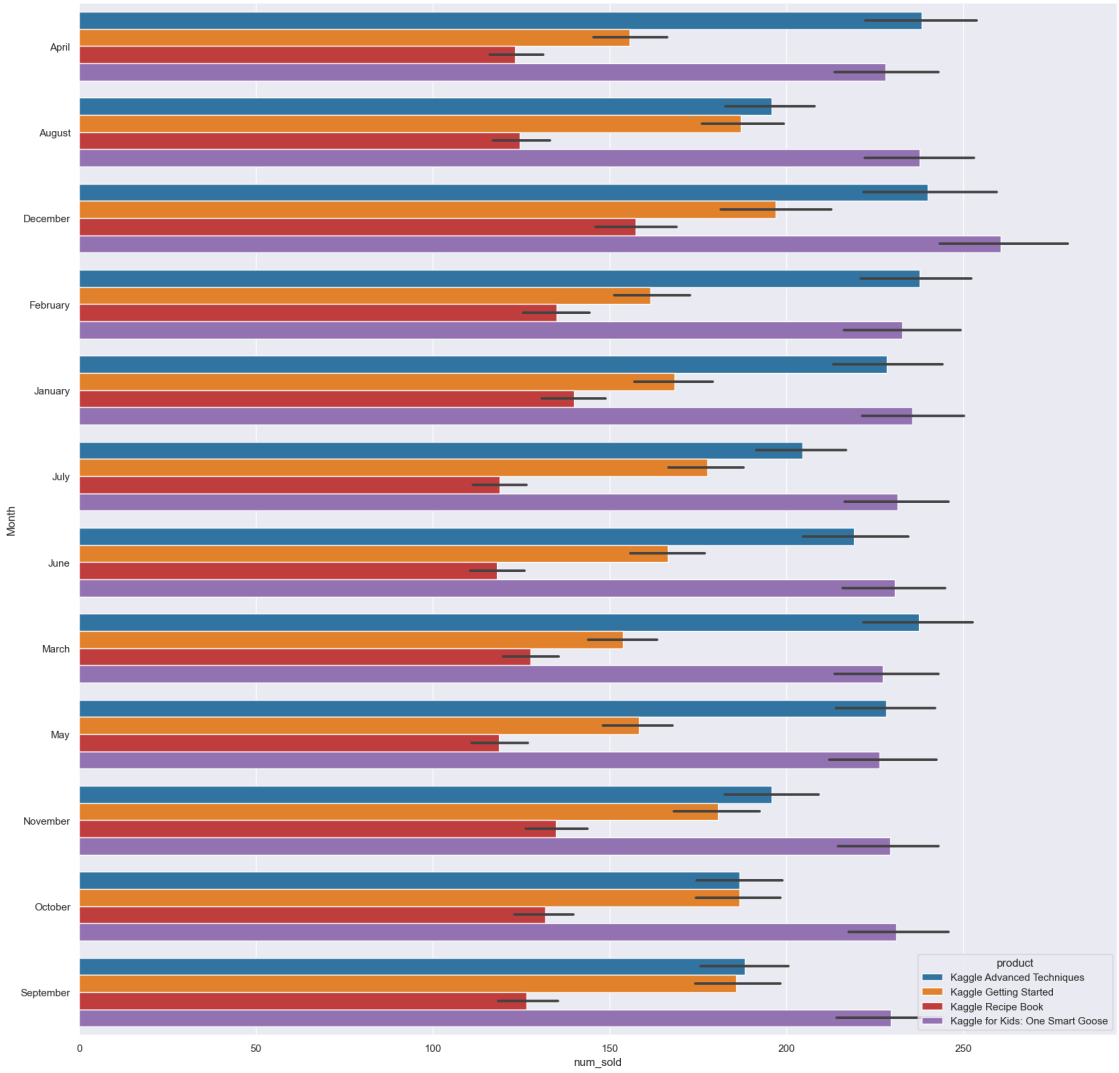
[89]:
```
plt.figure(figsize=(20,20))
sns.set_theme(style="darkgrid")
sns.barplot(x='num_sold', y='Date', data=sum9, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'],␣
↪orient= 'h')
```

[89]:
```
<AxesSubplot: xlabel='num_sold', ylabel='Date'>
```



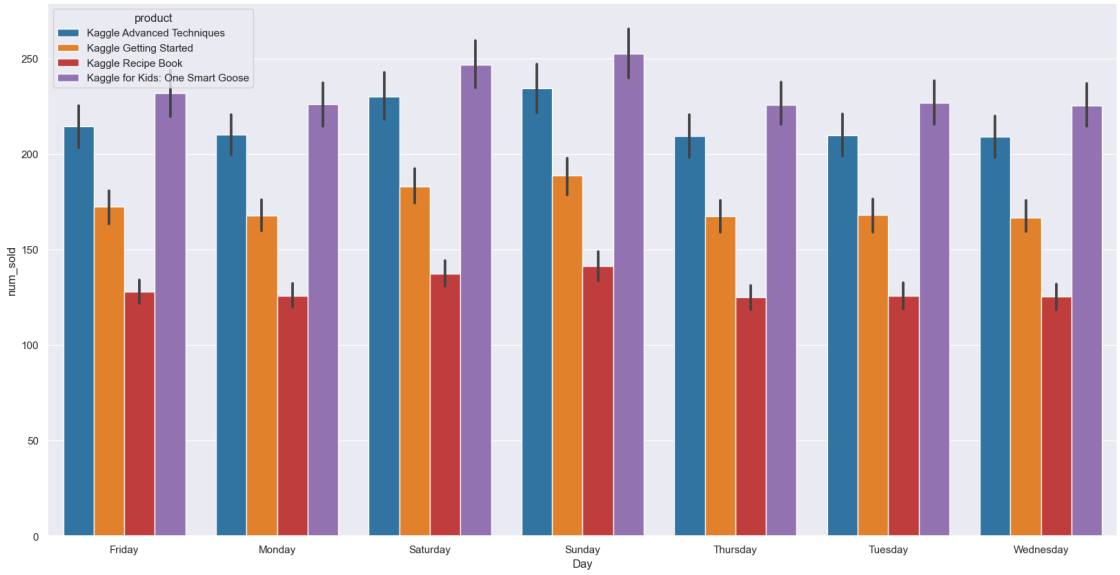
[91]:
```
plt.figure(figsize=(20,20))
sns.barplot(x='num_sold', y='Month', data=sum9, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'],␣
↪orient= 'h')
```

[91]:
```
<AxesSubplot: xlabel='num_sold', ylabel='Month'>
```



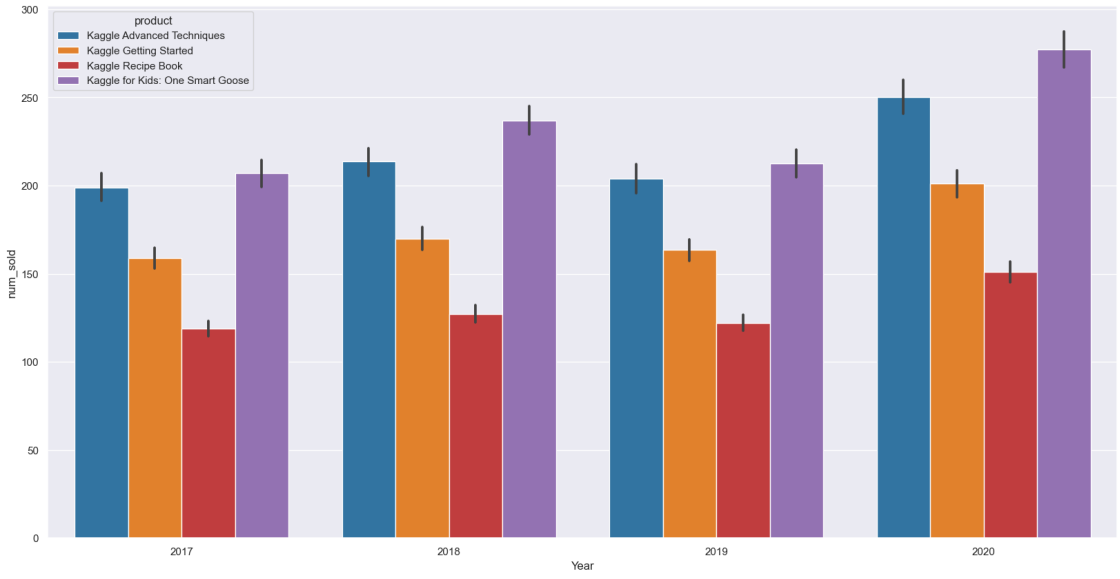
[92]:
```
plt.figure(figsize=(20,10))
sns.barplot(y='num_sold', x='Day', data=sum9, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'])
```

[92]:
```
<AxesSubplot: xlabel='Day', ylabel='num_sold'>
```



[82]:
```
plt.figure(figsize=(20,10))
sns.barplot(y='num_sold', x='Year', data=sum9, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'])
```

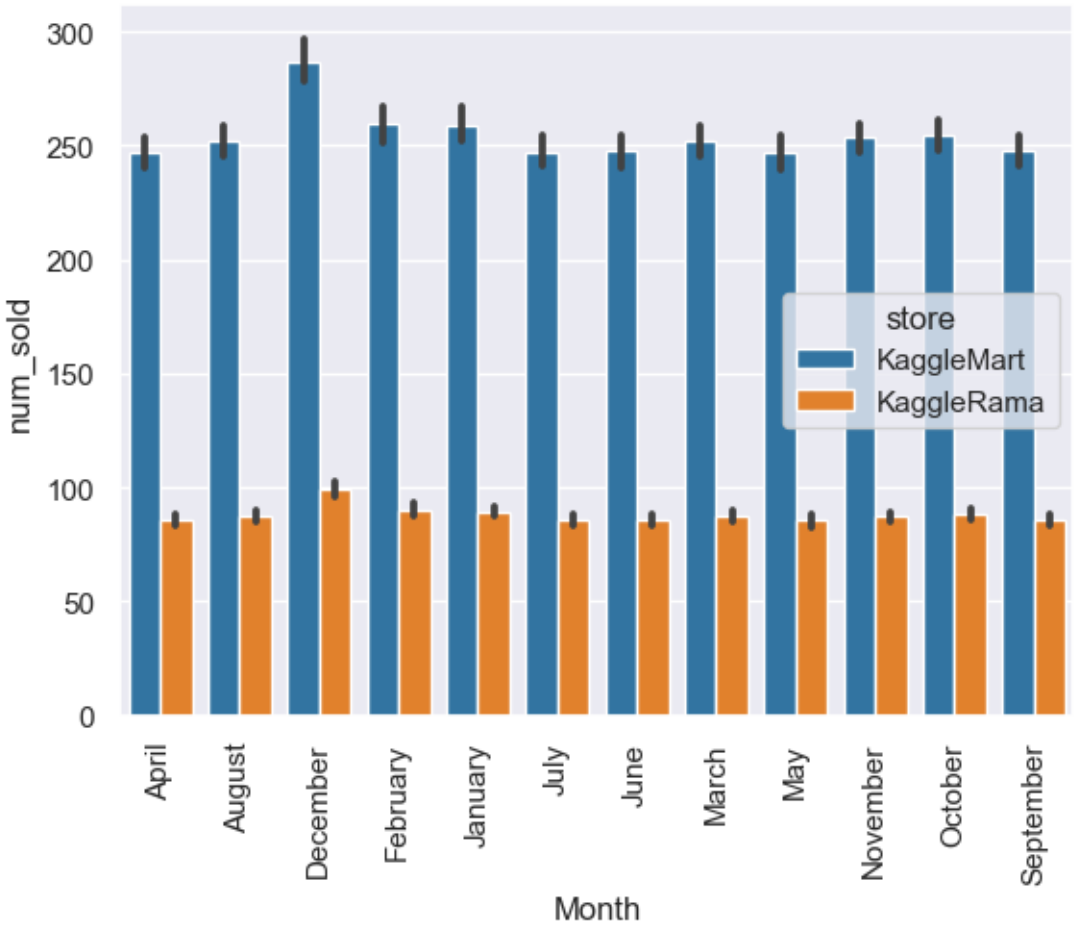
[82]: <AxesSubplot: xlabel='Year', ylabel='num_sold'>



0.6 Spain

[95]:
```
Spain = data[data.country == 'Spain']
sum10 = Spain.
 ↪groupby(['Month','store','Day','Year','Date','product'])['num_sold'].sum()
sum10 = pd.DataFrame(sum10)
sum10.reset_index(inplace=True)
sns.barplot(x='Month', y='num_sold', data=sum10, color='blue',␣
 ↪hue='store',palette = ['tab:blue', 'tab:orange'])
plt.xticks(rotation=90)
```

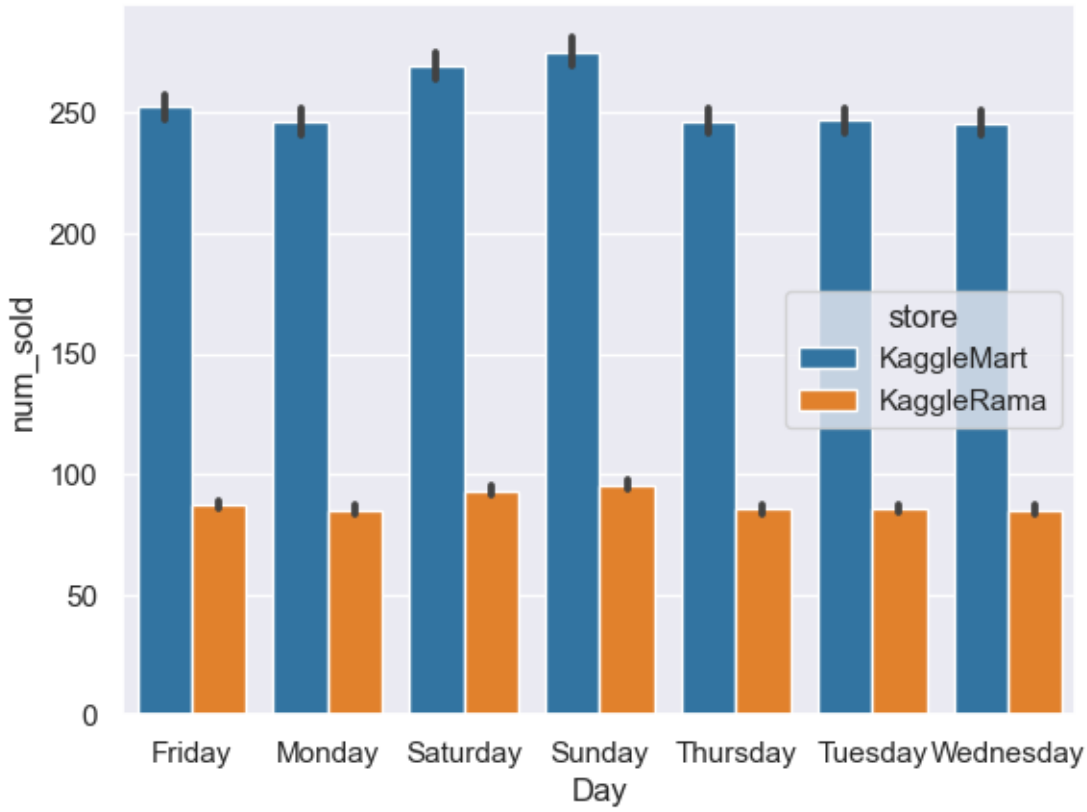
[95]:
```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11]),
 [Text(0, 0, 'April'),
  Text(1, 0, 'August'),
  Text(2, 0, 'December'),
  Text(3, 0, 'February'),
  Text(4, 0, 'January'),
  Text(5, 0, 'July'),
  Text(6, 0, 'June'),
  Text(7, 0, 'March'),
  Text(8, 0, 'May'),
  Text(9, 0, 'November'),
  Text(10, 0, 'October'),
  Text(11, 0, 'September')])
```



The most number of products in Spain are sold in December from the Kaggle Mart Store

[96]:
```
sns.barplot(x='Day', y='num_sold', data=sum10, color='blue',␣
↪hue='store',palette = ['tab:blue', 'tab:orange'])
```

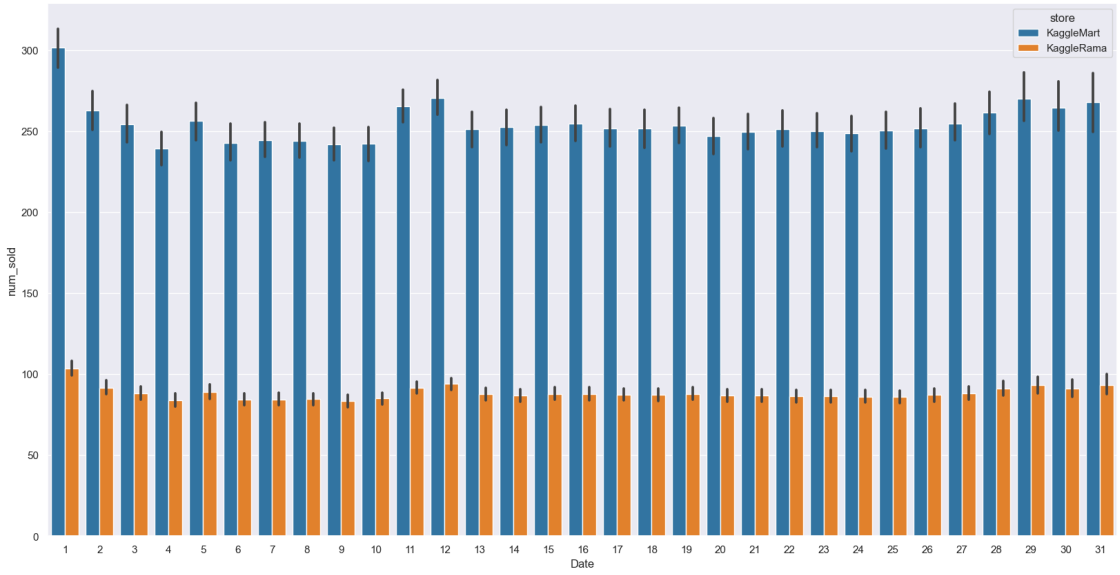
[96]:
```
<AxesSubplot: xlabel='Day', ylabel='num_sold'>
```



The most products are sold on Saturday and Sunday from Kaagle Mart Store in Spain

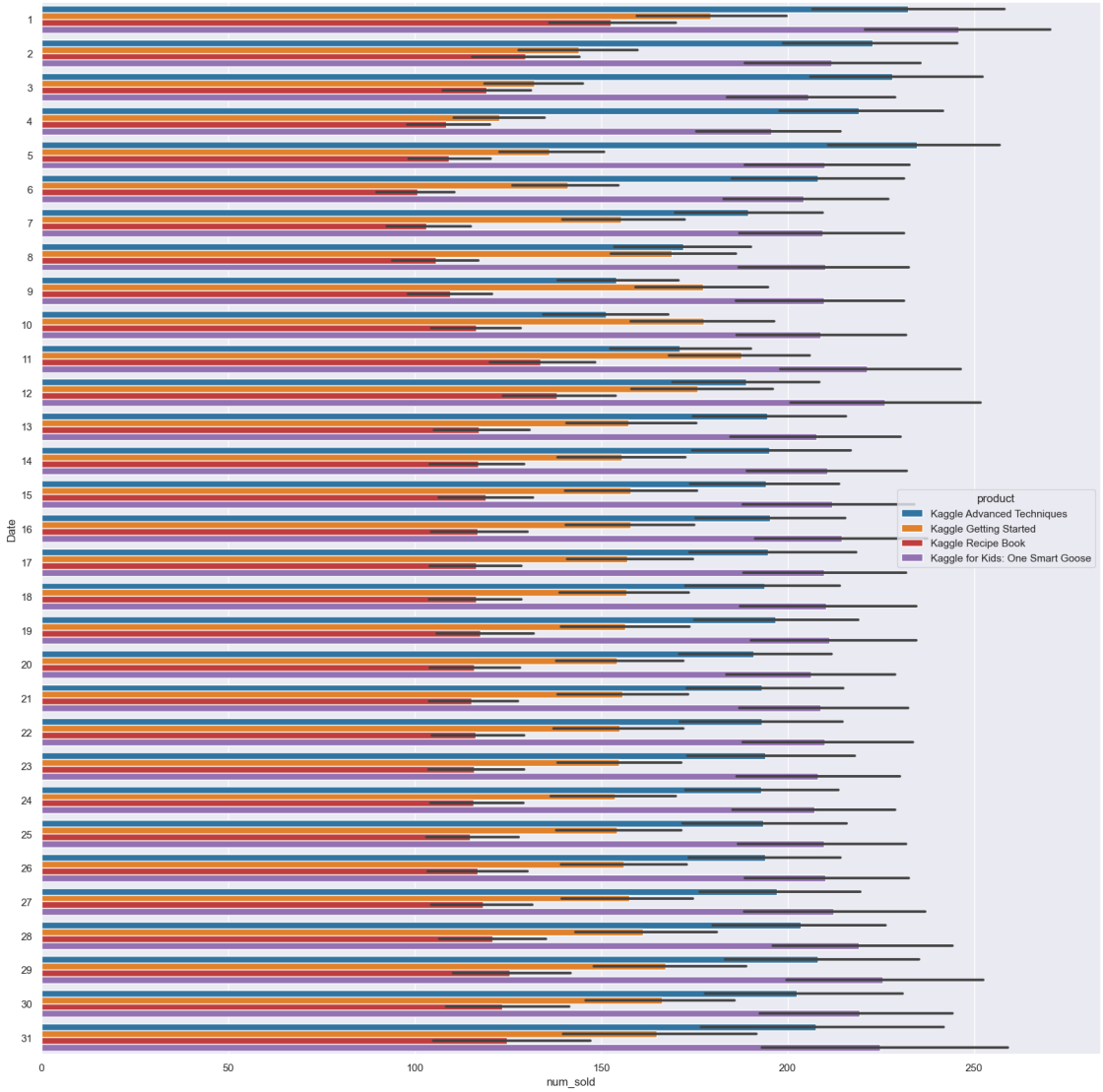
[97]:
```
plt.figure(figsize=(20,10))
sns.barplot(x='Date', y='num_sold', data=sum10, color='blue',␣
↪hue='store',palette = ['tab:blue', 'tab:orange'])
```

[97]:
```
<AxesSubplot: xlabel='Date', ylabel='num_sold'>
```



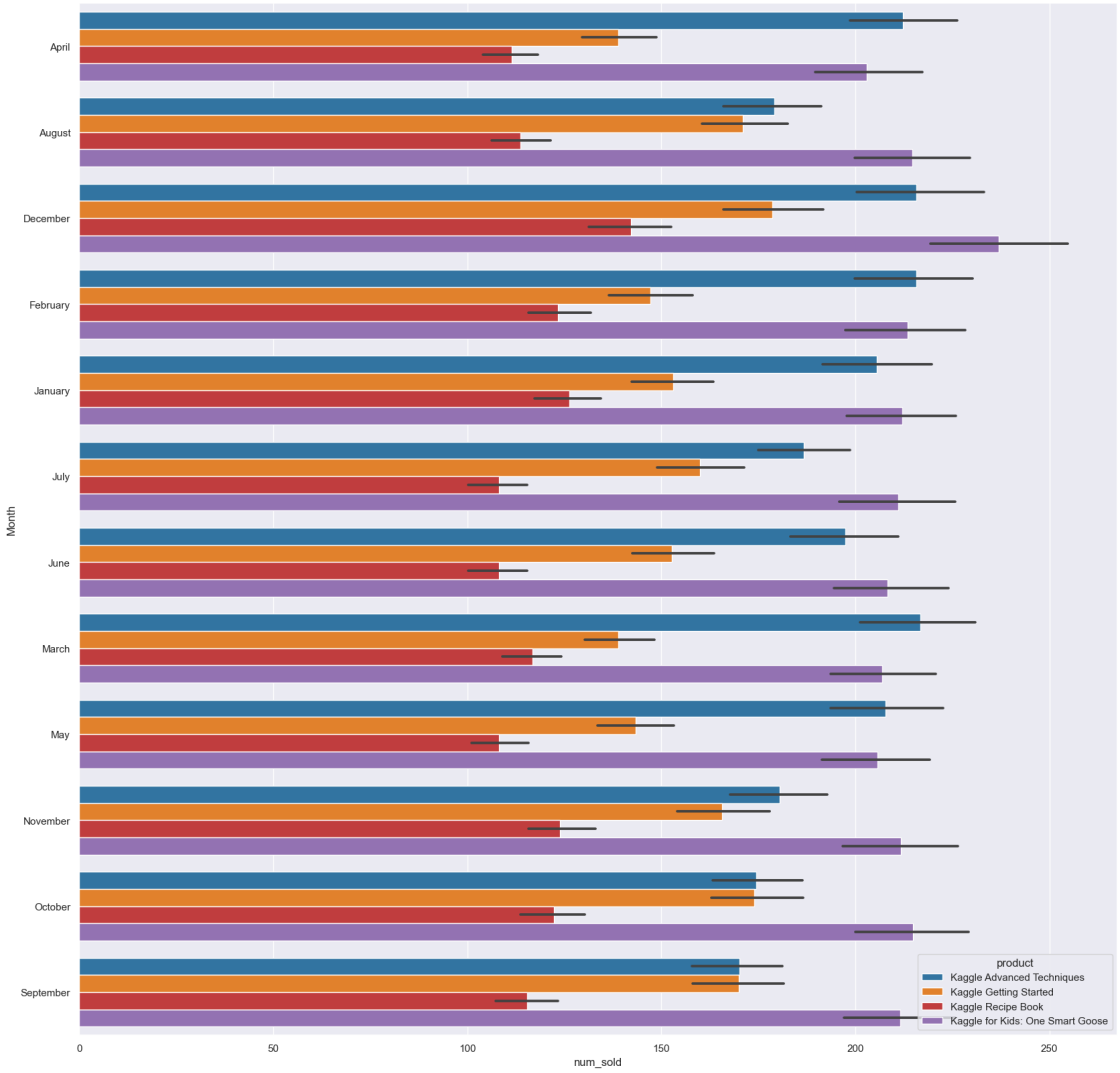
[98]:
```
plt.figure(figsize=(20,20))
sns.set_theme(style="darkgrid")
sns.barplot(x='num_sold', y='Date', data=sum10, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'],␣
↪orient= 'h')
```

[98]:
```
<AxesSubplot: xlabel='num_sold', ylabel='Date'>
```



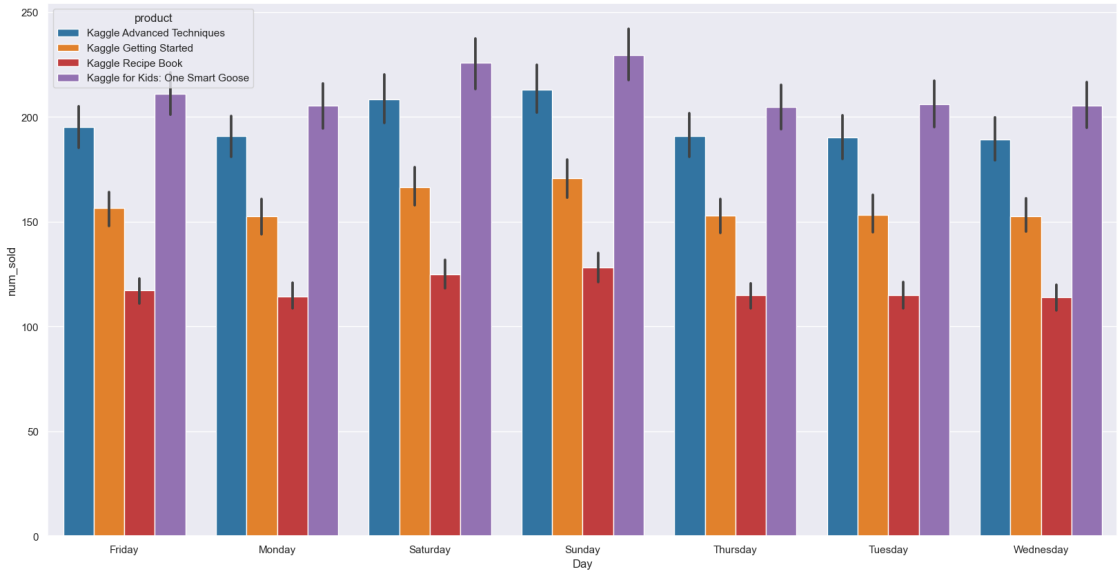
[99]:
```
plt.figure(figsize=(20,20))
sns.barplot(x='num_sold', y='Month', data=sum10, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'],␣
↪orient= 'h')
```

[99]:
```
<AxesSubplot: xlabel='num_sold', ylabel='Month'>
```



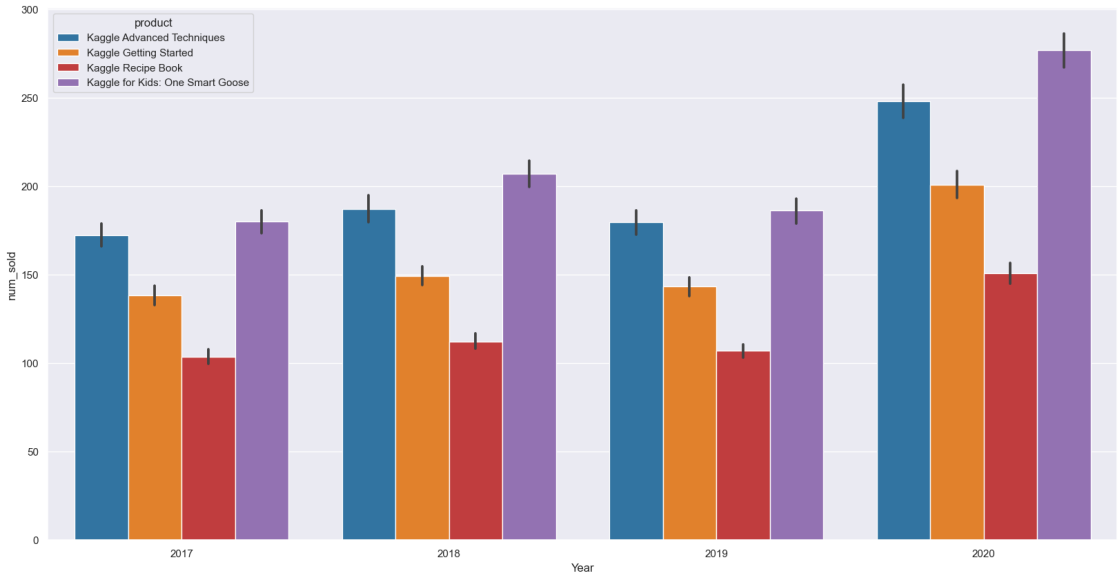
[100]:
```
plt.figure(figsize=(20,10))
sns.barplot(y='num_sold', x='Day', data=sum10, color='blue',␣
 ↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'])
```

[100]:
```
<AxesSubplot: xlabel='Day', ylabel='num_sold'>
```



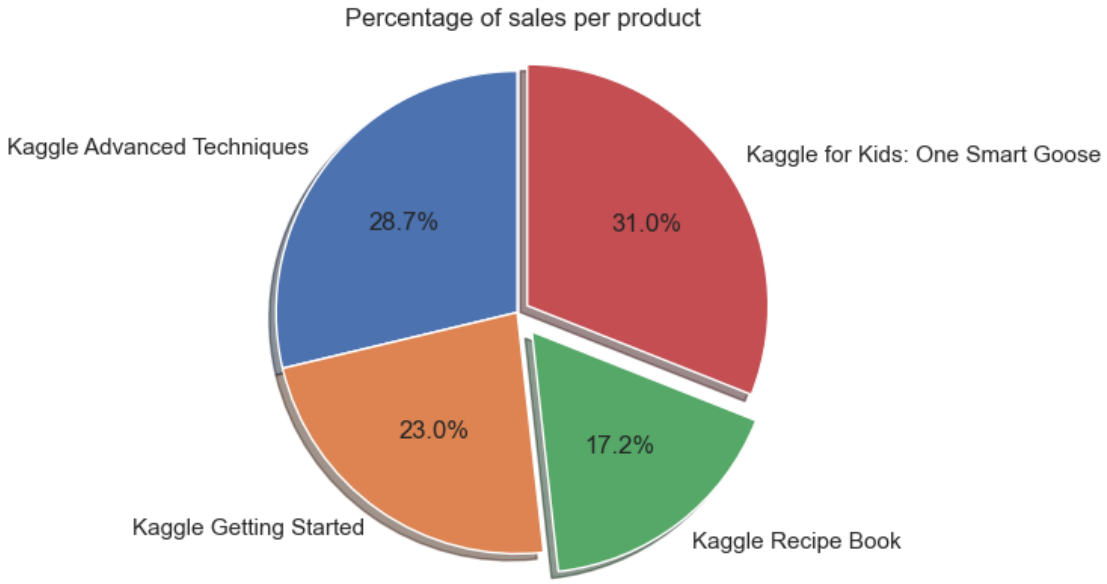
[101]:
```
plt.figure(figsize=(20,10))
sns.barplot(y='num_sold', x='Year', data=sum10, color='blue',␣
↪hue='product',palette = ['tab:blue', 'tab:orange','tab:red', 'tab:purple'])
```

[101]: <AxesSubplot: xlabel='Year', ylabel='num_sold'>



0.7 OverAll Analysis

[167]:
```
labels = data['product'].value_counts().index
sizes = [data.groupby('product')['num_sold'].sum()[0], data.
 ↪groupby('product')['num_sold'].sum()[1], data.groupby('product')['num_sold'].
 ↪sum()[2], data.groupby('product')['num_sold'].sum()[3]]
explode = (0, 0, 0.1, 0.05)
fig1, ax1 = plt.subplots()
ax1.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%',
        shadow=True, startangle=90)
ax1.axis('equal')
plt.title('Percentage of sales per product')
plt.show()
```



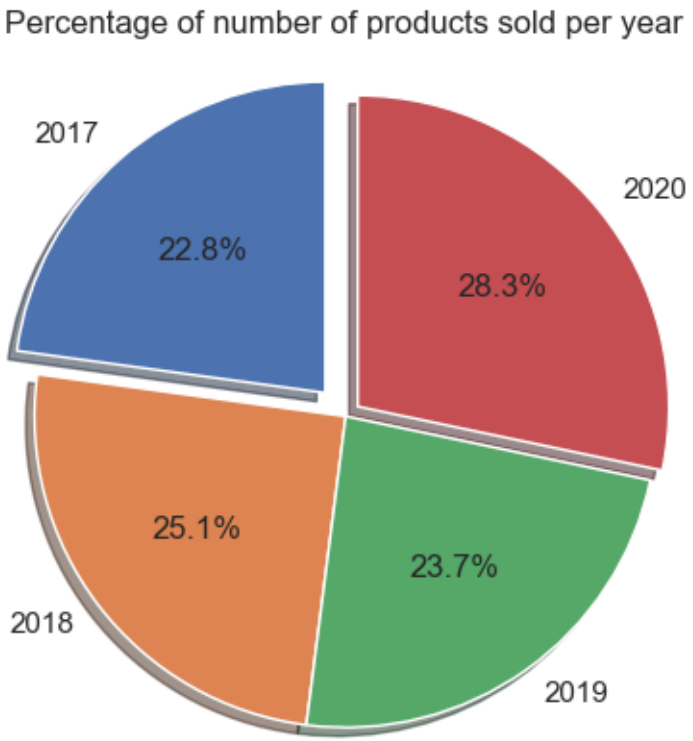
The lowest percentage of products sold is Kaggle Recipe Book and the most sold product is Kaggle for Kids: One Smart Goose of the total products sold

[152]:
```
one = pd.DataFrame(data.groupby('Year')['num_sold'].sum())
one.reset_index(inplace=True)
one
```

[152]:

	Year	num_sold
0	2017	3112163
1	2018	3425424
2	2019	3232879
3	2020	3855193

[154]:
```
labels = one.Year.value_counts().index
sizes = [one.num_sold[0],one.num_sold[1],one.num_sold[2],one.num_sold[3]]
explode = (0.1, 0, 0, 0.05)
fig1, ax1 = plt.subplots()
ax1.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%',
        shadow=True, startangle=90)
ax1.axis('equal')
plt.title('Percentage of number of products sold per year')
plt.show()
```



The most percentage of products were sold in the year of 2020 and th least were sold in 2017

[155]:
```
two = pd.DataFrame(data.groupby('Day')['num_sold'].sum())
two.reset_index(inplace=True)
two
```

[155]:

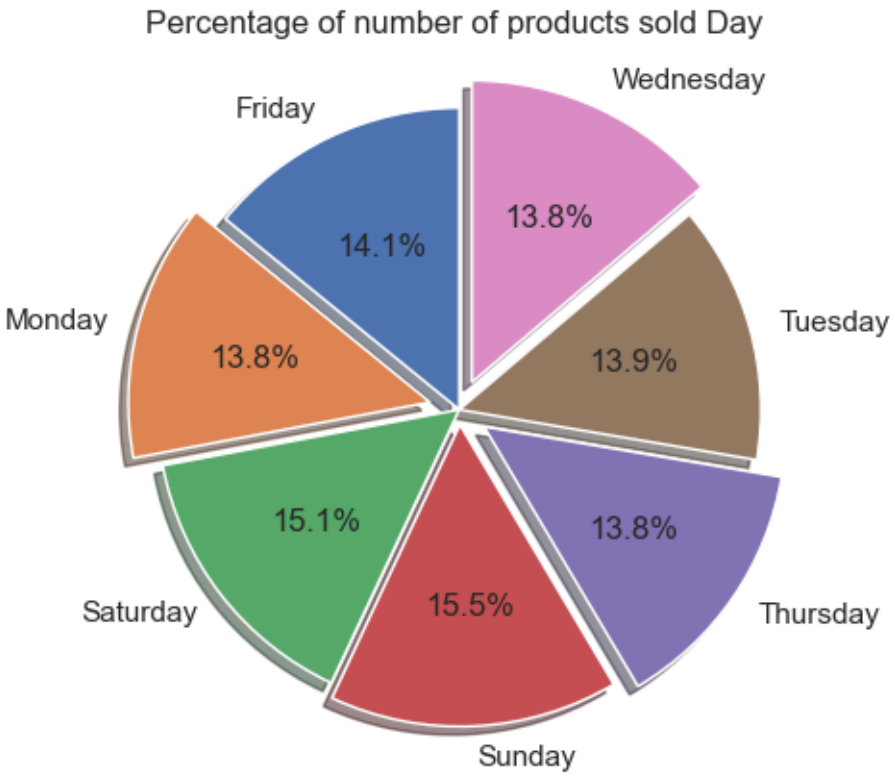
	Day	num_sold
0	Friday	1923624
1	Monday	1886783
2	Saturday	2052394
3	Sunday	2110734
4	Thursday	1882318

```
5    Tuesday   1889098
6  Wednesday   1880708
```

[162]:
```
two.Day.value_counts().index
```

[162]:
```
Index(['Friday', 'Monday', 'Saturday', 'Sunday', 'Thursday', 'Tuesday',
       'Wednesday'],
      dtype='object')
```

[166]:
```
labels = two.Day.value_counts().index
sizes = [two.num_sold[0],two.num_sold[1],two.num_sold[2],two.num_sold[3],two.
↪num_sold[4],two.num_sold[5],two.num_sold[6]]
explode = (0, 0.1, 0, 0.05, 0.1, 0, 0.1)
fig1, ax1 = plt.subplots()
ax1.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%',
        shadow=True, startangle=90)
ax1.axis('equal')
plt.title('Percentage of number of products sold Day')
plt.show()
```



Maximum ratio of products are sold on Sundays and Saturdays and least ratio of products are sold on Monday, Wednesdays and Thursday

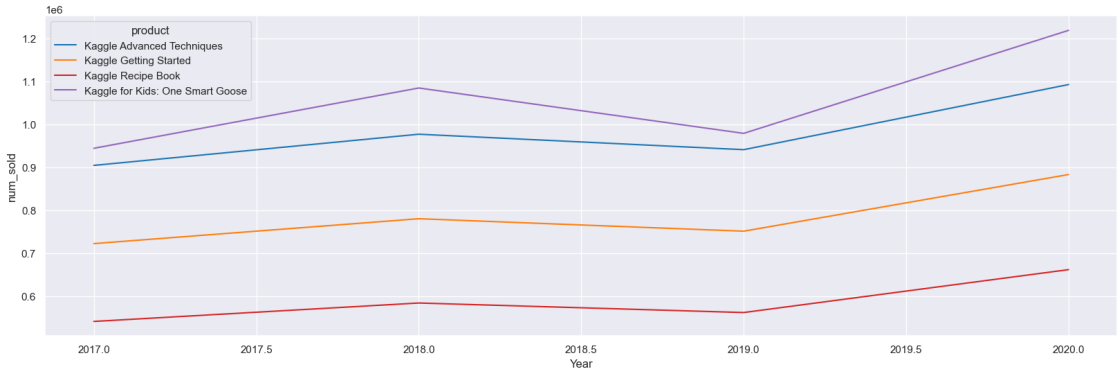
[191]:
```
three = pd.DataFrame(data.groupby(['Year','product'])['num_sold'].sum())
three.reset_index(inplace=True)
three.head()
```

[191]:

	Year	product	num_sold
0	2017	Kaggle Advanced Techniques	904253
1	2017	Kaggle Getting Started	722422
2	2017	Kaggle Recipe Book	541478
3	2017	Kaggle for Kids: One Smart Goose	944010
4	2018	Kaggle Advanced Techniques	976711

[177]:
```
plt.figure(figsize=(20,6))
sns.lineplot(x='Year', y='num_sold', data=three, hue='product', palette = ['tab:
↪blue', 'tab:orange','tab:red', 'tab:purple'])
```

[177]: <AxesSubplot: xlabel='Year', ylabel='num_sold'>



The sales of all the product dropped down in start of 2019 but it eventually rose high in the mids of 2019 and 2020

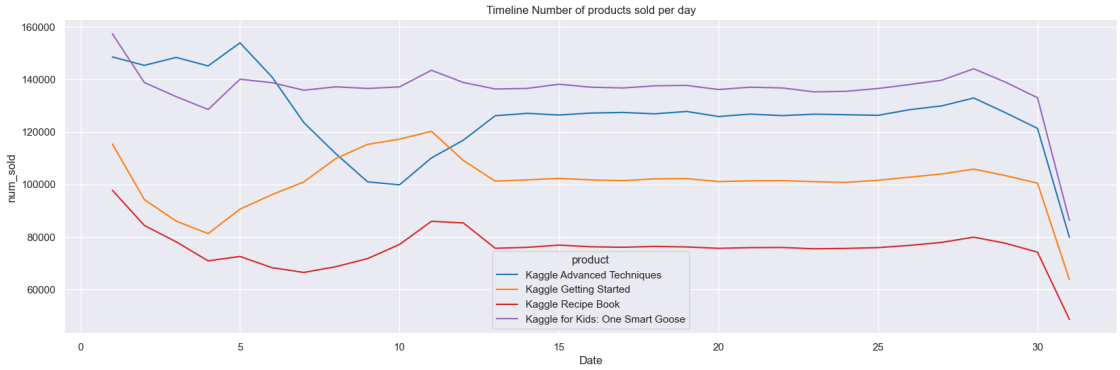
[190]:
```
four = pd.DataFrame(data.groupby(['Date','product'])['num_sold'].sum())
four.reset_index(inplace=True)
four.head()
```

[190]:

	Date	product	num_sold
0	1	Kaggle Advanced Techniques	148621
1	1	Kaggle Getting Started	115413
2	1	Kaggle Recipe Book	97774
3	1	Kaggle for Kids: One Smart Goose	157432
4	2	Kaggle Advanced Techniques	145418

[183]:
```
plt.figure(figsize=(20,6))
sns.lineplot(x='Date', y='num_sold', data=four, hue='product', palette = ['tab:
↪blue', 'tab:orange','tab:red', 'tab:purple'])
```

```
plt.title('Timeline Number of products sold per day')
plt.show()
```

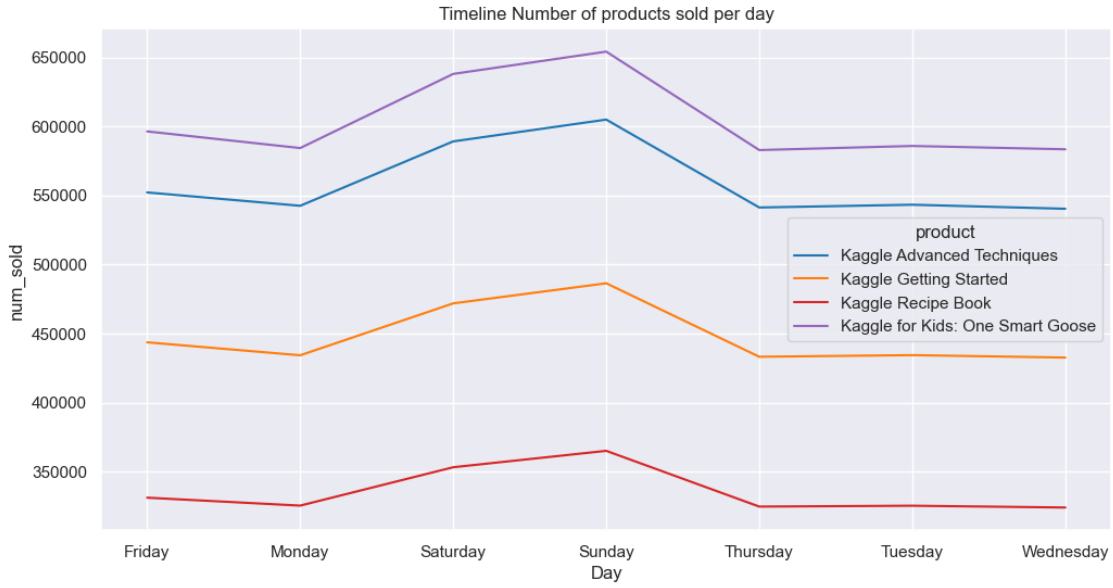


[189]:
```
five = pd.DataFrame(data.groupby(['Day','product'])['num_sold'].sum())
five.reset_index(inplace=True)
five.head()
```

[189]:

	Day	product	num_sold
0	Friday	Kaggle Advanced Techniques	552219
1	Friday	Kaggle Getting Started	443726
2	Friday	Kaggle Recipe Book	331289
3	Friday	Kaggle for Kids: One Smart Goose	596390
4	Monday	Kaggle Advanced Techniques	542552

[188]:
```
plt.figure(figsize=(12,6))
sns.lineplot(x='Day', y='num_sold', data=five, hue='product', palette = ['tab:
 ↪blue', 'tab:orange','tab:red', 'tab:purple'])
plt.title('Timeline Number of products sold per day')
plt.show()
```



Products sales rises at Saturday and Sunday

CONCLUSION

- The following are the conclusion drawn after performing Exploratory Data Analysis:

1. Most number of the sales were done in the year of 2020.
2. Most number of products are sold were in the month of December.
3. Most number of sales were done on Saturday's and Sunday's which were 15.1% and 15.9% respectively.
4. Largest number of products sold were "Kaggle for Kids: One Smart Goose".
5. Most number of products were sold in Belgium.
6. Most orders of a product were received on the date of 30-12-2020.