FINDING THE BEST MARKETS TO ADVERTISE IN

In 2017, an online learning platform called FreeCodeCamp that offers courses in web development, Mobile development, data science and game development wanted to advertise its courses. To know what courses to advertise and which region to target, FreeCodeCamp surveyed their students to base their advertising decisions on first-hand data from their students.

The responses to this survey were later posted to FreeCodeCamp's github repository https://github.com/freeCodeCamp/2017-new-coder-survey.

The survey had more than 130 questions, some about students' interests, their educational background, and other questions to determine whether students will afford courses they were going to be advertised to; and it was filled by more than 18,000 students from all over the world.

We'll going to look into these students' responses and try to answer the two main questions that FreeCodeCamp needed to answer:

1. What course/courses should they advertise; and
2. what region should they target

IMPORTING AND EXPLORING THE DATASET

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
online_df = 'https://raw.githubusercontent.com/freeCodeCamp/2017-new-coder-survey/master/clean-data/2017-fCC-New-C
dataset = pd.read_csv(online_df, low_memory = 0)
```

In [2]:
```
dataset.shape
```

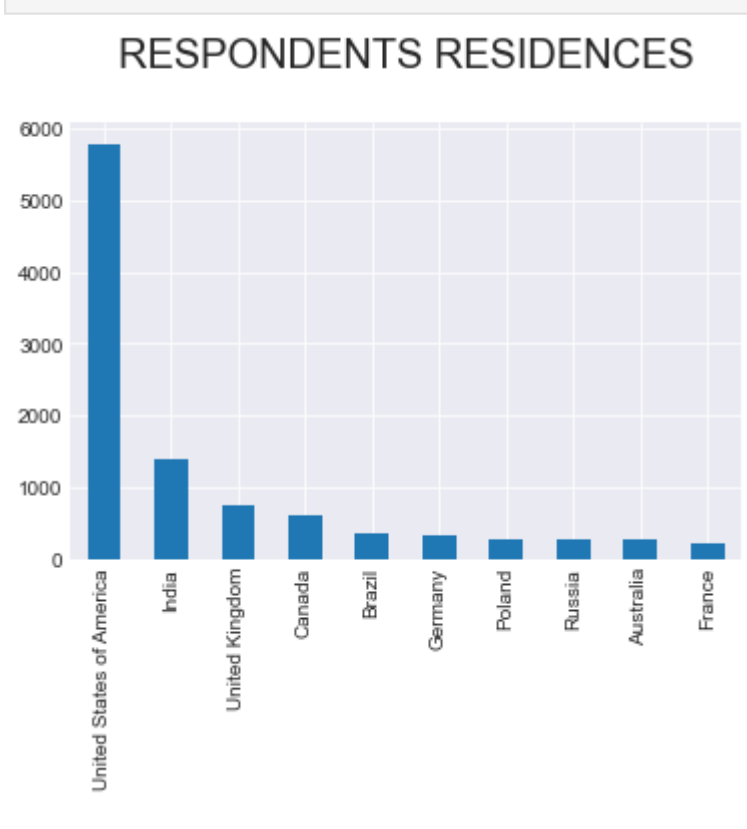
Out[2]:
```
(18175, 136)
```

In [3]:
```
pd.options.display.max_columns = 140
dataset.head()
```

Out[3]:

	Age	AttendedBootcamp	BootcampFinish	BootcampLoanYesNo	BootcampName	BootcampRecommend	ChildrenNumber	CityPopulation	Coc
0	27.0	0.0	NaN	NaN	NaN	NaN	NaN	more than 1 million	
1	34.0	0.0	NaN	NaN	NaN	NaN	NaN	less than 100,000	
2	21.0	0.0	NaN	NaN	NaN	NaN	NaN	more than 1 million	
3	26.0	0.0	NaN	NaN	NaN	NaN	NaN	between 100,000 and 1 million	
4	20.0	0.0	NaN	NaN	NaN	NaN	NaN	between 100,000 and 1 million	

In [4]:
```
%matplotlib inline
plt.style.use('seaborn-darkgrid')
residences = dataset['CountryLive'].value_counts()[:10].plot.bar().title
plt.title("RESPONDENTS RESIDENCES", fontsize = 20, y = 1.10)
plt.show()
```



The majority of students who filled the survey were from the US as shown by the above chart. While this is an important detail, we still need to know how many of those respondents indicated courses they would like to take, and whether they can afford the 59 usd monthly subscription.

NUMBER OF COURSE INTERESTS PER STUDENT

In [23]:
```
cleaned_interests = dataset['JobRoleInterest'].dropna()
split = cleaned_interests.str.split(',')
number_of_interests = split.apply(lambda x: len(x))
(number_of_interests.value_counts(normalize = True) * 100)
```

Out[23]:
```
1     31.650458
3     15.889588
4     15.217391
5     12.042334
2     10.883867
6      6.721968
7      3.861556
8      1.759153
9      0.986842
10     0.471968
12     0.300343
11     0.185927
13     0.028604
Name: JobRoleInterest, dtype: float64
```

Around 32% of respondents have already narrowed their interests to one specific course they would like to take, whereas the other 68% have wide interests ranging from 3 to 13 different courses.

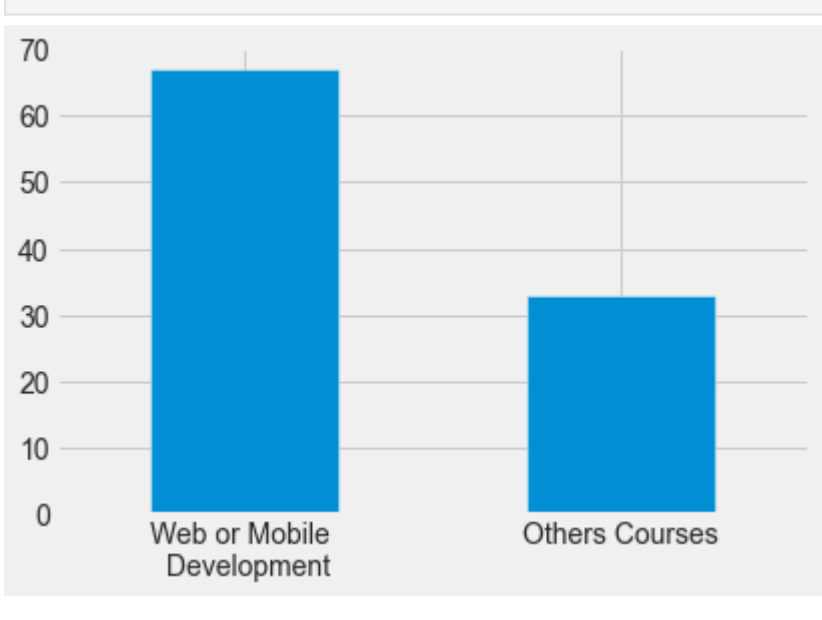
2 possible decisions can be made from this insight:

1. Focus the advertisment towards the 32% who knows what course or career track they want to pursue - as they are likely to invest in developing their skills than the ones who are still thinking or exploring the right course to invest the time and resources in.
2. Alternatively, offering the other 68% who haven't narrowed their interests a free trial might convert them into subscribers after they explore different courses and decide on which one to invest in.

WHAT COURSES ARE STUDENTS INTERESTED IN

FreeCodeCamp offers a lot of courses with the main ones being mobile development and web development. We would like to know if these are still the main interests of the students, or if other courses should be developed more and prioritized when it comes time to advertise.

In [24]:
```
web_mobile_int = cleaned_interests.str.contains("Mobile Developer|Web De veloper")
freq_web_mobile = web_mobile_int.value_counts(normalize = True)*100
plt.style.use("fivethirtyeight")
freq_web_mobile.plot.bar()
plt.xticks([0,1], ['Web or Mobile \n Development', 'Others Courses'], rotation = 0)
plt.show()
```



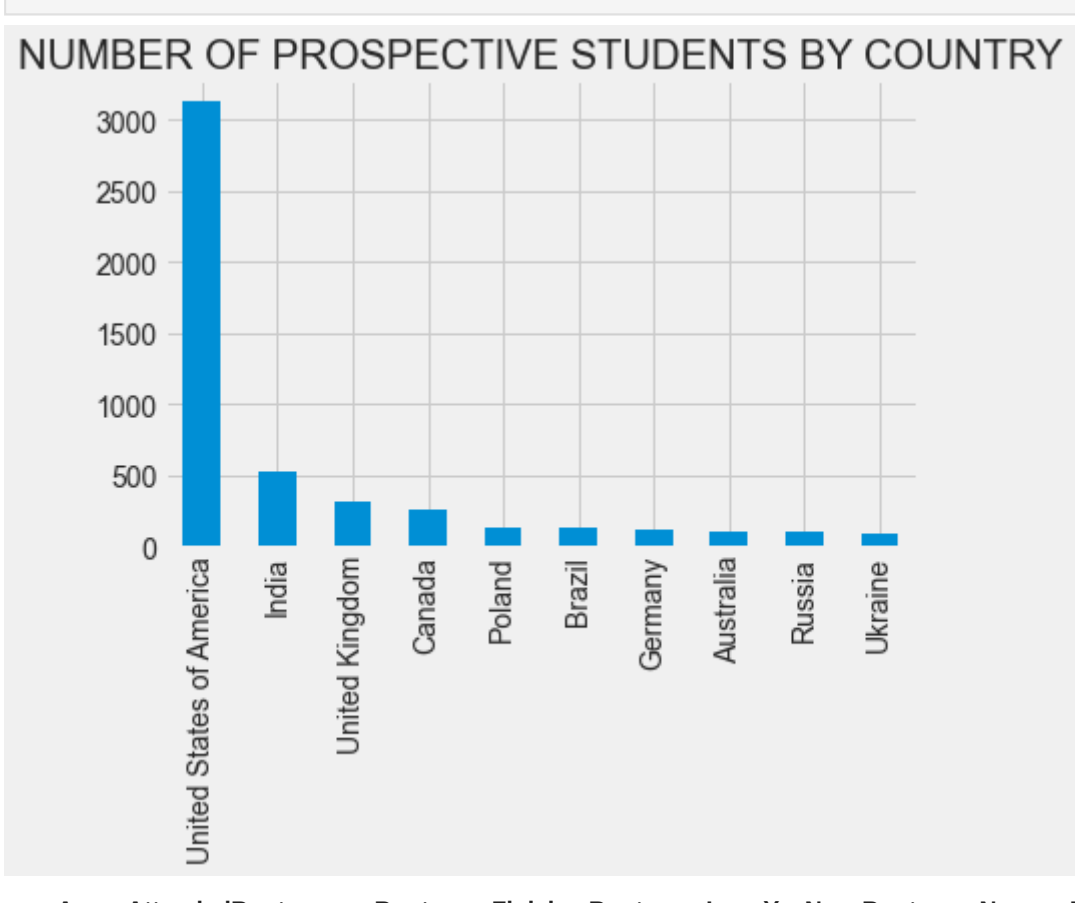
As the above graph shows, web and mobile development are still the main interests of students with more than 60 % of students indicating a desire to take courses in the 2 areas.

Based on this, these 2 courses should be prioritized when advertising.

But, in what geographical areas should be advertised in first?

BIGGEST MARKETS BY NUMBER OF STUDENTS

In [9]:
```
cleaned = dataset[dataset['JobRoleInterest'].notnull()]
cleaned['CountryLive'].value_counts().sort_values(ascending = False)[:10].plot.bar()
plt.style.use('fivethirtyeight')
plt.title("NUMBER OF PROSPECTIVE STUDENTS BY COUNTRY")
plt.show()
cleaned.head()
```



Out[9]:

	Age	AttendedBootcamp	BootcampFinish	BootcampLoanYesNo	BootcampName	BootcampRecommend	ChildrenNumber	CityPopulation	Coc
1	34.0	0.0	NaN	NaN	NaN	NaN	NaN	less than 100,000	
2	21.0	0.0	NaN	NaN	NaN	NaN	NaN	more than 1 million	
3	26.0	0.0	NaN	NaN	NaN	NaN	NaN	between 100,000 and 1 million	
4	20.0	0.0	NaN	NaN	NaN	NaN	NaN	between 100,000 and 1 million	
6	29.0	0.0	NaN	NaN	NaN	NaN	NaN	between 100,000 and 1 million	

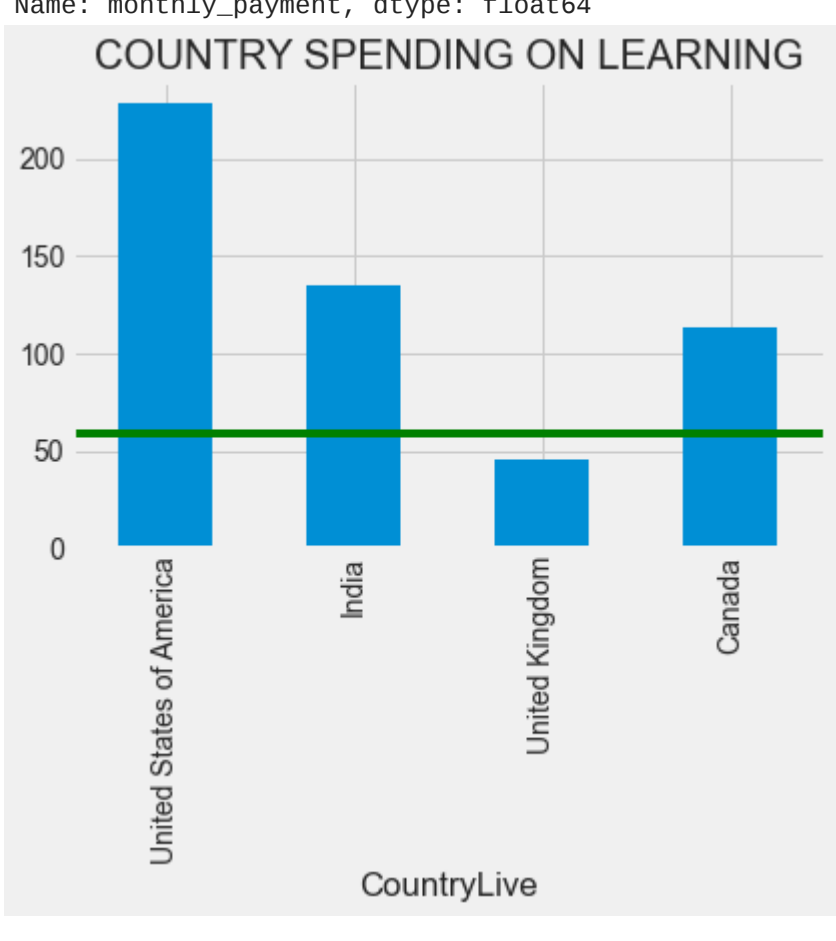
The US has the most students who indicated their job interests. The next biggest are India, UK and Canada.

Next, we'll examine whether our monthly subscription is affordable for these students. We'll also look at how much they usually spend on their learning per month compared to our 59 usd monthly subscription.

STUDENTS PURCHASING POWER BY COUNTRY

In [10]:
```
pd.options.mode.chained_assignment = None
cleaned['MonthsProgramming'].replace(0,1,inplace = True)
cleaned['monthly_payment'] = (cleaned['MoneyForLearning'])/(cleaned['MonthsProgramming'])
cleaned[cleaned['monthly_payment'].notnull()]
cleaned = cleaned[cleaned['CountryLive'].notnull()]
cleaned['MonthsProgramming']
country_grouped = cleaned.groupby('CountryLive').mean()
afford = country_grouped['monthly_payment'][['United States of America',
'India','United Kingdom','Canada']]
print(afford)
afford.plot.bar()
plt.axhline(y =59, color = 'green', label = '$59 Monthly Subscription')
plt.title("COUNTRY SPENDING ON LEARNING")
plt.show()
```

```
CountryLive
United States of America    227.997996
India                       135.100982
United Kingdom               45.534443
Canada                      113.510961
Name: monthly_payment, dtype: float64
```



Respondents from the US, India and Canada usually spend more than FreeCodeCamp monthly subscription of 59 usd on their learning. Those from the UK, however, indicated to spend less than 50 usd monthly.

We suspect that these respondents from the UK who indicated to have lower purchasing power than their counterparts from India might not be represetative of the whole UK market. With the UK GDP per capita being around 20x that of India, it would be unlikely for students in India to have a purchasing power 3 times more that of UK students.

We also suspect that we have significant outliers that are biasing our current analysis:

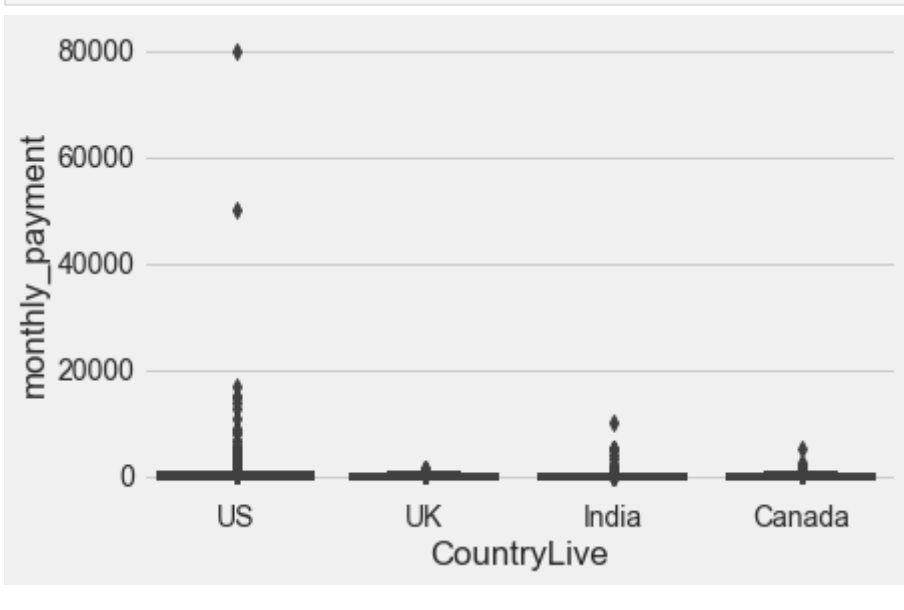
- Some respondents might have included their college tuition which should not have been included in the amount they spent on learning - the survey only asks what students spent on their learning other than their college tuition. In this case, responses for the 'MoneyForLearning' column will be inflated as well as those of 'Monthly_Payment' that is derived from 'MoneyForLearning'.
- Some respondents might have joined Tech Bootcamps whose tuition will also inflate the 'MoneyForLearning'. This will raise their average 'Monthly_Payment' significantly since most bootcamps charge tens of thousands of US dollars and their programs are completed within 3-6 months.
- Other respondents might have used free learning resources while supplementing them with few paid courses as needed. This will significantly lower the 'Monthly_Payment' values for those respondents since they have been studying for a long time and have only paid for few resources

Clearly there are a lot of possible scenarios that will bias our analysis.

In the next part, we'll going to remove outliers in our dataset if their responses seem extremely unusual, and then we'll confirm what is the best market to advertise in.

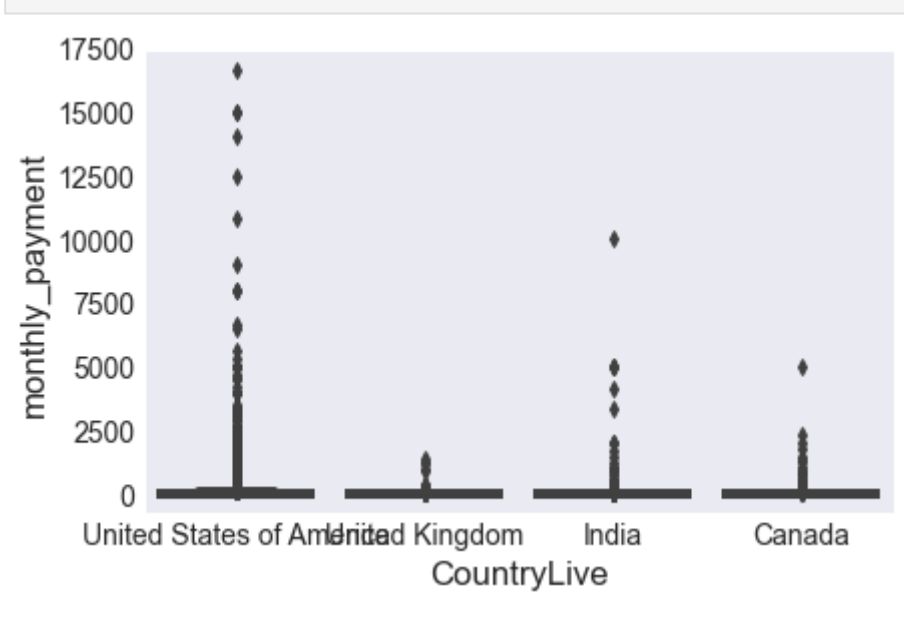
ADJUSTING FOR OUTLIERS

In [25]:
```
import warnings
warnings.filterwarnings('ignore')
big_markets = cleaned[cleaned['CountryLive'].str.contains('United States of America|India|United Kingdom|Canada')]
sns.boxplot(data = big_markets, y ='monthly_payment', x = 'CountryLive')
sns.set_style('dark')
sns.color_palette('deep')
plt.xticks(range(4), ['US', 'UK', 'India', 'Canada'])
plt.show()
```



One of the respondents who lives in the US, as shown by the above box plots, indicated to invest more than 50,000 usd per month. Since this is most likely to be a mistake either from data entry mistake or from misunderstanding of one of the question on the survey, we'll drop this data point.

In [26]:
```
big_markets = big_markets[big_markets['monthly_payment']<50000]
sns.boxplot(data = big_markets, x = 'CountryLive', y ='monthly_payment')
plt.show()
```



In [27]:
```
 usa_outliers = big_markets[(big_markets['CountryLive']=='United States of America') & (big_markets['monthly_paymen
usa_outliers['AttendedBootcamp'].value_counts()
```

Out[27]:
```
1.0    6
0.0    5
Name: AttendedBootcamp, dtype: int64
```

In [28]:
```
usa_outliers
```

Out[28]:

	Age	AttendedBootcamp	BootcampFinish	BootcampLoanYesNo	BootcampName	BootcampRecommend	ChildrenNumber	CityPopulation
718	26.0	1.0	0.0	0.0	The Coding Boot Camp at UCLA Extension	1.0	NaN	more than 1 million
1222	32.0	1.0	0.0	0.0	The Iron Yard	1.0	NaN	between 100,000 and 1 million
3184	34.0	1.0	1.0	0.0	We Can Code IT	1.0	NaN	more than 1 million
3930	31.0	0.0	NaN	NaN	NaN	NaN	NaN	between 100,000 and 1 million
6805	46.0	1.0	1.0	1.0	Sabio.la	0.0	NaN	between 100,000 and 1 million
7198	32.0	0.0	NaN	NaN	NaN	NaN	NaN	more than 1 million
7505	26.0	1.0	0.0	1.0	Codeup	0.0	NaN	more than 1 million
9778	33.0	1.0	0.0	1.0	Grand Circus	1.0	NaN	between 100,000 and 1 million
16650	29.0	0.0	NaN	NaN	NaN	NaN	2.0	more than 1 million
16997	27.0	0.0	NaN	NaN	NaN	NaN	1.0	more than 1 million
17231	50.0	0.0	NaN	NaN	NaN	NaN	2.0	less than 100,000

6 of 11 outliers in US datapoints went to bootcamps which justify their average monthly spending on learning because bootcamps charge tens of thousands of dollards for programs that are completed within few months.

For the other 5 outliers, nothing in the dataset indicate why they might have spent over 6000 usd per month learning. We think there might have been a data entry mistake or some misunderstandings that inflated these numbers, and so we'll drop these data points

In [29]:
```
big_markets = big_markets.drop(usa_outliers.index)
india_outliers= big_markets[(big_markets['monthly_payment']>4000) &(big_markets['CountryLive']=='India')]

india_outliers
```

Out[29]:

	Age	AttendedBootcamp	BootcampFinish	BootcampLoanYesNo	BootcampName	BootcampRecommend	ChildrenNumber	CityPopulation
1728	24.0	0.0	NaN	NaN	NaN	NaN	NaN	between 100,000 and 1 million
7989	28.0	0.0	NaN	NaN	NaN	NaN	NaN	between 100,000 and 1 million
8126	22.0	0.0	NaN	NaN	NaN	NaN	NaN	more than 1 million
13398	19.0	0.0	NaN	NaN	NaN	NaN	NaN	more than 1 million
15587	27.0	0.0	NaN	NaN	NaN	NaN	NaN	more than 1 million

We'll also drop the five outliers from India bceause nothing in the dataset indicates why they would have spent 5000 usd, 10,000 usd and even 100,000 usd per month on their learning.

In [30]:
```
big_markets = big_markets.drop(india_outliers.index)
```

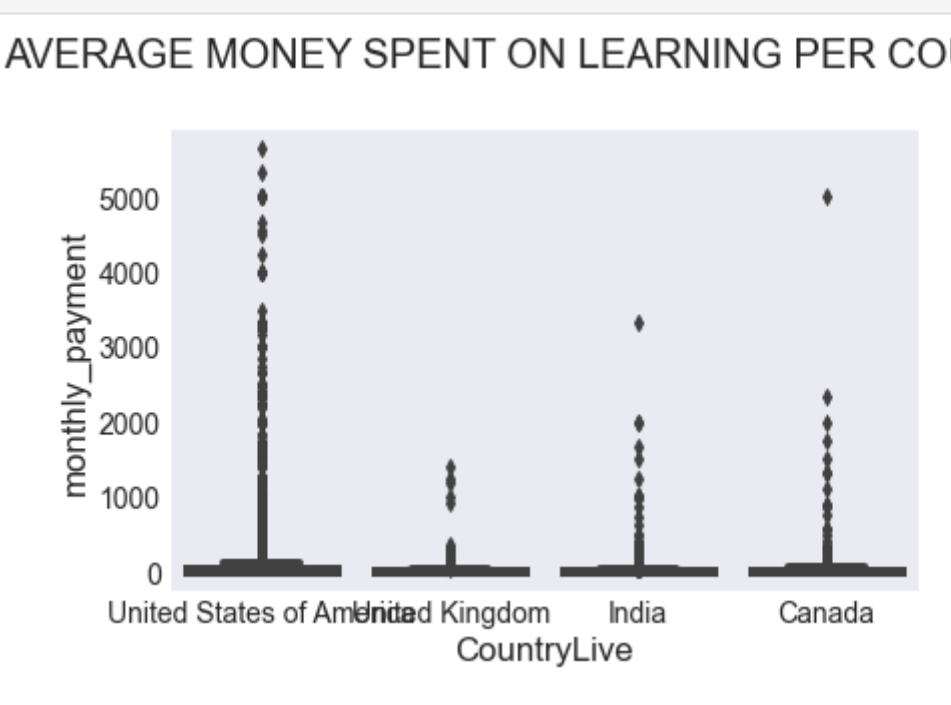
In [31]:
```
canada_outlier = big_markets[(big_markets['CountryLive']=='Canada') &(big_markets['monthly_payment']>4000)]
canada_outlier
```

Out[31]:

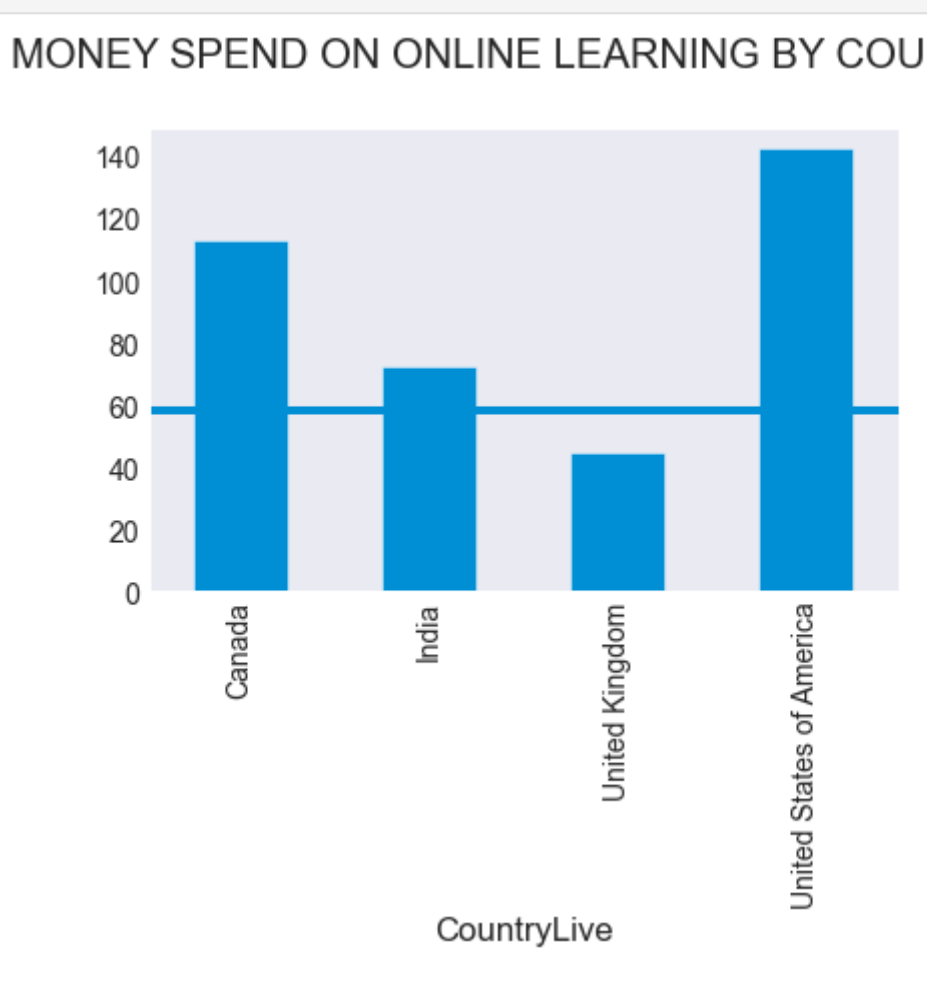
	Age	AttendedBootcamp	BootcampFinish	BootcampLoanYesNo	BootcampName	BootcampRecommend	ChildrenNumber	CityPopulation
13659	24.0	1.0	0.0	0.0	Bloc.io	1.0	NaN	more than 1 million

The only outlier from respondents who reside in Canada was enrolled in a bootcamp when he/she filled the survey and so we'll keep this data point in our analysis.

In [32]:
```
sns.boxplot(data = big_markets, x = 'CountryLive', y = 'monthly_payment')
plt.title("AVERAGE MONEY SPENT ON LEARNING PER COUNTRY", fontsize = 20,y =1.1)
plt.show()
```



In [33]:
```
biggest= big_markets.groupby('CountryLive')
biggest['monthly_payment'].mean().plot.bar()
plt.title("MONEY SPEND ON ONLINE LEARNING BY COUNTRY", y = 1.1)
plt.axhline(y=59)
plt.show()
```



In [34]:
```
biggest['monthly_payment'].mean()
```

Out[34]:
```
CountryLive
Canada                      113.510961
India                        72.893205
United Kingdom               45.534443
United States of America    142.654608
Name: monthly_payment, dtype: float64
```

CONCLUSION

The best market for FreeCodeCamp to advertise in is the US, for two main reasons:

The majority of students interested in learning Web Development, Mobile Development and Data Science live in the US.

And the average monthly spending on learning for US tech students of 142 usd far exceeds our 59 usd monthly subscription as shown in the analysis.

To conclude, we recommend to prioritize advertising in the US if FreeCodeCamp want to start their advertisment campaign in one country. But if they have enough budget to advertise in all big markets, we recommend spending around 70% of that budget in the US, 15% in Canada and 15% in India.