In [1]:

```
import pandas as pd
import numpy as nm
from matplotlib import pyplot as plt
```

In [2]:

```
%%html

<div class="text-center">
<h1> Evaluation of Stack Overflow Survey</h1>
</div>

<h2>Data: </h2>
        <p> Review of over 65,000 respondents to a survey <br>
        </p>

        <h3>From Stack Overflow:<br></h3>
        "This year, we focused on seeking diverse representation while asking for inf
        from technologies and behavior to questions that will help us improve the Sta
        for everybody who codes."
        </p>

<h2> Aggregation and review of data below by the following metrics:<br></h2>
    <p>
        &#9642; Reported salary in US dollars<br>
        &#9642; Respondents by Country <br>
        &#9642; Programming languages currently known <br>
        &#9642; Programming languages desired to learn<br><br>

        <p> </br>


    survey source: https://insights.stackoverflow.com/survey/2020
    </p>
```

Evaluation of Stack Overflow Survey

Data:

Review of over 65,000 respondents to a survey

From Stack Overflow:

"This year, we focused on seeking diverse representation while asking for information ranging from technologies and behavior to questions that will help us improve the Stack Overflow community for everybody who codes."

Aggregation and review of data below by the following metrics:

- Reported salary in US dollars
- Respondents by Country
- Programming languages currently known
- Programming languages desired to learn

survey source: https://insights.stackoverflow.com/survey/2020

In [3]:
```
df = pd.read_csv('2020Data/survey_results_public.csv', index_col='Respondent')
schema_df = pd.read_csv('2020Data/survey_results_schema.csv', index_col='Column')
```

In [4]:
```
pd.set_option('display.max_columns', 60)
pd.set_option('display.max_rows', 85)
pd.set_option('display.float_format', lambda x: '%.3f' % x)
```

In [5]:
```
df.rename(columns={'ConvertedComp': 'SalaryUSD'}, inplace=True)
```

In [6]:
```
#drop inaccurate result
df.drop(14419, inplace=True)
```

In [7]:
```
#Filter Data
filt = df['SalaryUSD'] < 2000000
```

General Salary and Age Information Listed Below

- After sorting total respondands are 30,332
- Mean Salary is $89,441 median salary of $54,049
- Mean Age is 32, median age is 30

In [28]:
```
df[['Country','SalaryUSD','Age']].median()
```

Out[28]:
```
SalaryUSD   54049.000
Age            30.000
dtype: float64
```

In [29]:
```
df[['Country','SalaryUSD','Age']].describe()
```

Out[29]:

	SalaryUSD	Age
count	30332.000	30332.000
mean	89441.404	32.100
std	156068.844	8.319
min	0.000	15.000
25%	25915.750	26.000
50%	54049.000	30.000
75%	93533.000	36.000
max	1980000.000	69.000

Per the graphs and data frame below:

- The survey in general is skewed towards younger respondents. This is to be expected since the survey is conducted from Stack Overflow and the average age of a software developer is 32 years old. (The average working age across all professions is 42.3)

- Subsequent graphs note the median salary (USD) for the top five countries responding to the survey.
US Salary is notably above the world median. India is below except for one outlier.

- In general, all data follows the expected trend that the greater the age, and likely experience, the higher the salary. This trend does drop off after nearing 60 years old.

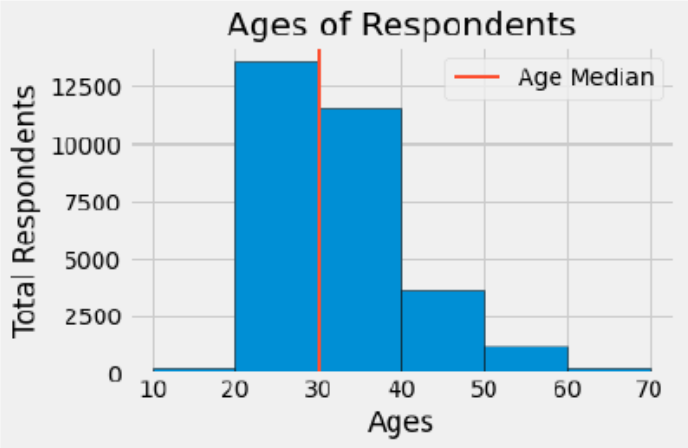
Median Salary and Age by Country

In [35]:
```
median_df
```

Out[35]:

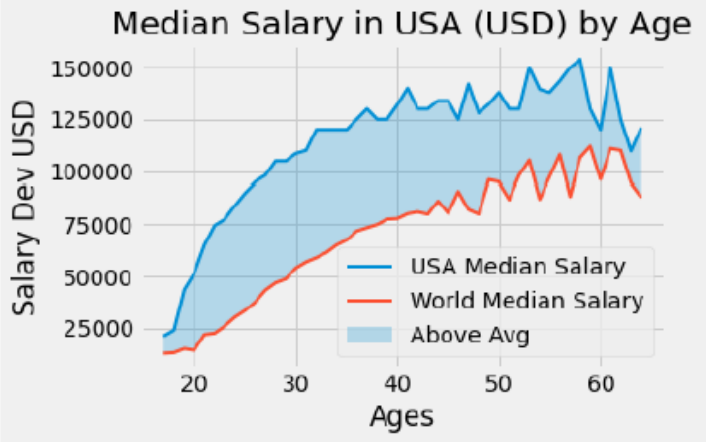
	Median Salary USD	Median Age
All	54049.000	30.000
United States	111000.000	32.000
United Kingdom	67215.000	31.000
Canada	68068.000	31.000
Germany	62697.000	31.000
India	10471.000	26.000

In [36]:
```
plt.style.use('fivethirtyeight')
bins = [10, 20, 30, 40, 50, 60, 70]

plt.hist(filt_age, bins=bins, edgecolor='black', log=False)

color = '#fc4f30'
plt.axvline(median_age, color=color, label='Age Median', linewidth=2)

plt.title('Ages of Respondents')
plt.xlabel('Ages')
plt.ylabel('Total Respondents')
plt.tight_layout()
plt.legend()
plt.show()
```



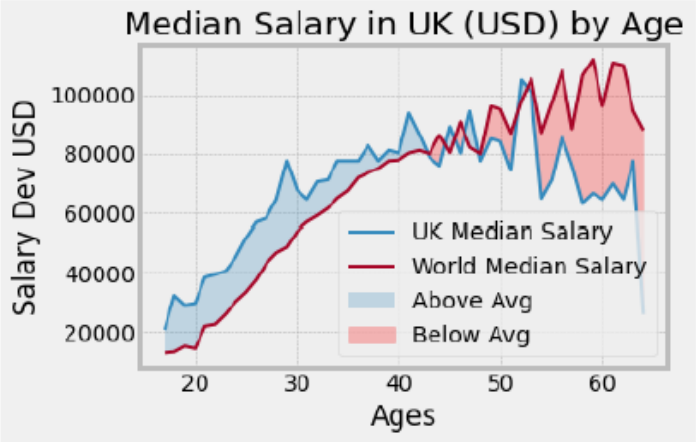
In [37]:

```
plt.xlabel('Ages')
plt.ylabel('Salary Dev USD')
plt.title('Median Salary in USA (USD) by Age')
plt.style.use('bmh')

plt.plot(age_All_x, salary_USA_y, label='USA Median Salary')
plt.plot(age_All_x, sal_All_y, label='World Median Salary')

plt.fill_between(age_All_x, salary_USA_y, sal_All_y,
                 where=(nm.asarray(salary_USA_y) > nm.asarray(sal_All_y)),
                 interpolate=True, alpha=0.25, label='Above Avg')

plt.fill_between(age_All_x, salary_USA_y, sal_All_y,
                 where=(nm.asarray(salary_USA_y) < nm.asarray(sal_All_y)),
                 interpolate=True, color='red', alpha=0.25)


plt.tight_layout()
plt.legend()
plt.show()
```



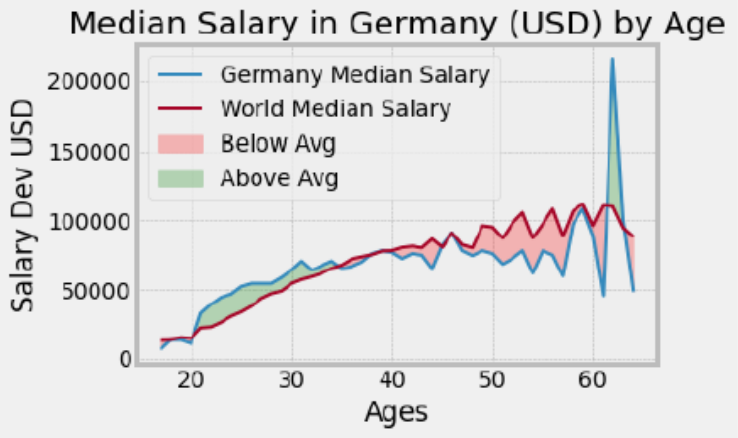
In [38]:

```
plt.xlabel('Ages')
plt.ylabel('Salary Dev USD')
plt.title('Median Salary in UK (USD) by Age')
plt.style.use('bmh')

plt.plot(age_All_x, salary_UK_y, label='UK Median Salary')
plt.plot(age_All_x, sal_All_y, label='World Median Salary')

plt.fill_between(age_All_x, salary_UK_y, sal_All_y,
                 where=(nm.asarray(salary_UK_y) > nm.asarray(sal_All_y)),
                 interpolate=True, alpha=0.25, label='Above Avg')

plt.fill_between(age_All_x, salary_UK_y, sal_All_y,
                 where=(nm.asarray(salary_UK_y) < nm.asarray(sal_All_y)),
                 interpolate=True, color='red', alpha=0.25, label='Below Avg')


plt.tight_layout()
plt.legend()
plt.show()
```



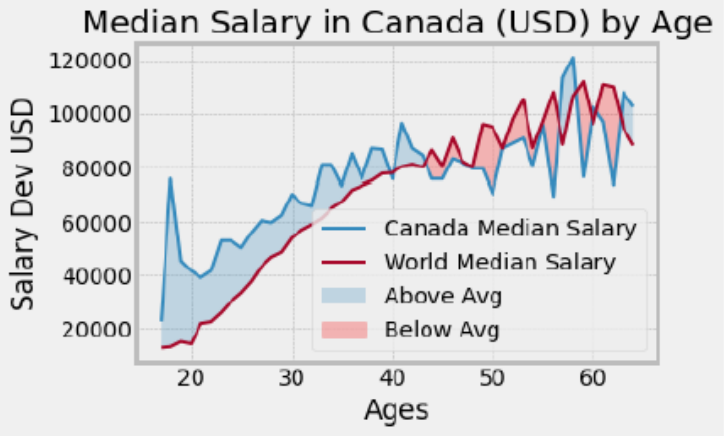
In [39]:

```
plt.xlabel('Ages')
plt.ylabel('Salary Dev USD')
plt.title('Median Salary in Germany (USD) by Age')
plt.style.use('bmh')

plt.plot(age_All_x, salary_Germ_y, label='Germany Median Salary')
plt.plot(age_All_x, sal_All_y, label='World Median Salary')

plt.fill_between(age_All_x, salary_Germ_y, sal_All_y,
                 where=(nm.asarray(salary_Germ_y) < nm.asarray(sal_All_y)),
                 interpolate=True, alpha=0.25, color='red', label='Below Avg')

plt.fill_between(age_All_x, salary_Germ_y, sal_All_y,
                 where=(nm.asarray(salary_Germ_y) >= nm.asarray(sal_All_y)),
                 interpolate=True, color='green', alpha=0.25, label='Above Avg')



plt.tight_layout()
plt.legend()
plt.show()
```



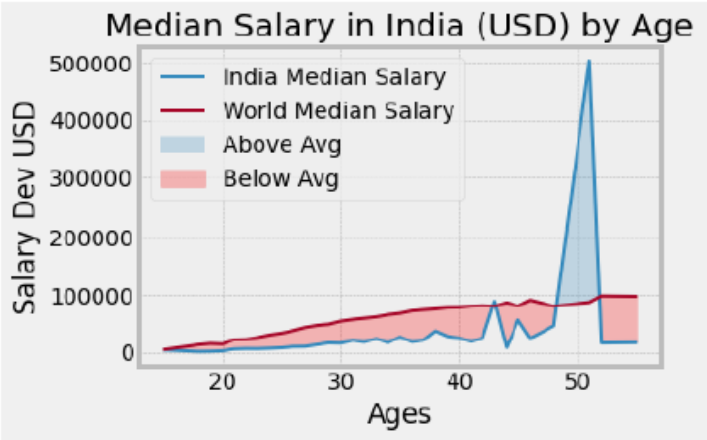
In [40]:

```
plt.xlabel('Ages')
plt.ylabel('Salary Dev USD')
plt.title('Median Salary in Canada (USD) by Age')
plt.style.use('bmh')

plt.plot(age_All_x, salary_Can_y, label='Canada Median Salary')
plt.plot(age_All_x, sal_All_y, label='World Median Salary')

plt.fill_between(age_All_x, salary_Can_y, sal_All_y,
                 where=(nm.asarray(salary_Can_y) > nm.asarray(sal_All_y)),
                 interpolate=True, alpha=0.25, label='Above Avg')

plt.fill_between(age_All_x, salary_Can_y, sal_All_y,
                 where=(nm.asarray(salary_Can_y) < nm.asarray(sal_All_y)),
                 interpolate=True, color='red', alpha=0.25, label='Below Avg')


plt.tight_layout()
plt.legend()
plt.show()
```



In [41]:

```
plt.xlabel('Ages')
plt.ylabel('Salary Dev USD')
plt.title('Median Salary in India (USD) by Age')
plt.style.use('bmh')

plt.plot(age_All_Ind_x, salary_Ind_y, label='India Median Salary')
plt.plot(age_All_Ind_x, sal_All_Ind_y, label='World Median Salary')

plt.fill_between(age_All_Ind_x, salary_Ind_y, sal_All_Ind_y,
                 where=(nm.asarray(salary_Ind_y) > nm.asarray(sal_All_Ind_y)),
                 interpolate=True, alpha=0.25, label='Above Avg')

plt.fill_between(age_All_Ind_x, salary_Ind_y, sal_All_Ind_y,
                 where=(nm.asarray(salary_Ind_y) < nm.asarray(sal_All_Ind_y)),
                 interpolate=True, color='red', alpha=0.25, label='Below Avg')


plt.tight_layout()
plt.legend()
plt.show()
```



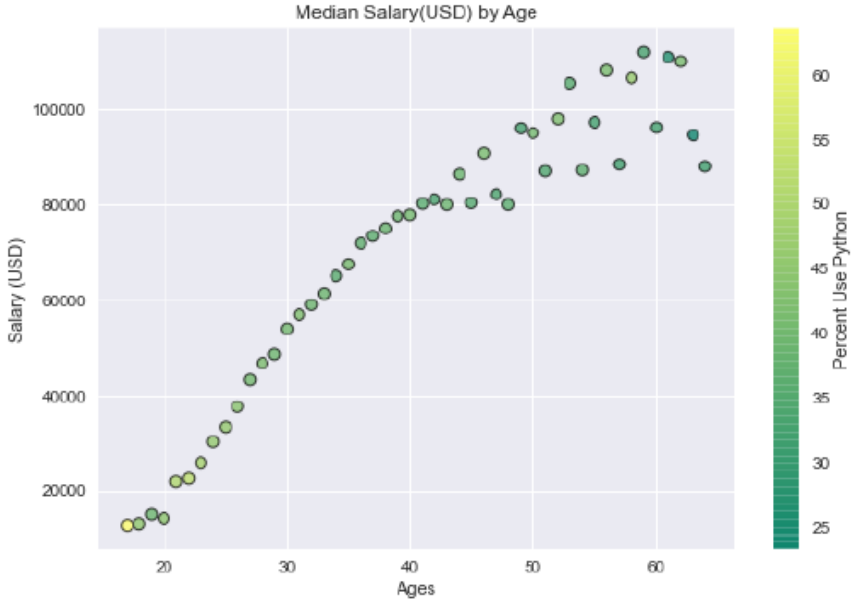
The graph below notes median salary (USD) for all countries and ages

•The color notes if respondents currently know Python.

(The more yellow the color, the higher percentage of respondents already know Python)

In [68]:

```
plt.style.use('seaborn')

plt.scatter(age_All_x, sal_All_y, c = Python_percent, cmap='summer',
            edgecolor='black', linewidth=1, alpha=0.75)

cbar = plt.colorbar()
cbar.set_label("Percent Use Python")

plt.title('Median Salary(USD) by Age')
plt.xlabel('Ages')
plt.ylabel('Salary (USD)')

plt.tight_layout()
plt.savefig('Sal_Age_Py.png')
```



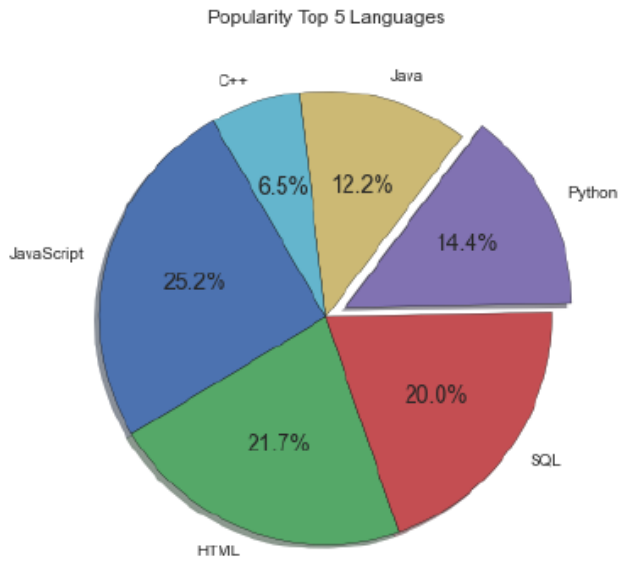
The pie charts below note the popularity of the top 5 programming languages

- The first chart shows languages currently known by respondents.
- The second chart shows languages currently respondents want to lear in 2021.

- These charts highlight the desire for more programmers to learn and use Python.
(Currently 14.4% of respondents know Python, but 21.3% of respondents want to learn Python in 2021)

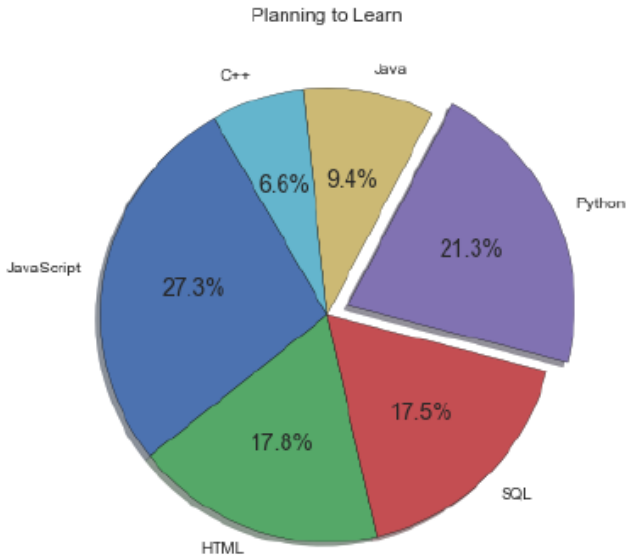
- The standard languages of JavaScript, HTML, and SQL remain popular. Judging from this information, Python looks to be the upcoming language most respondents want to learn in 2021.

In [75]:
```
labels = ['JavaScript', 'HTML', 'SQL', 'Python', 'Java', 'C++']
slices = [totalJavaScript, totalHTML, totalSQL, totalPy, totalJava, totalCpp]
explode = [0, 0, 0, 0.1, 0, 0]

plt.pie(slices, labels=labels, explode=explode, shadow=True, startangle=120, autopct=
       wedgeprops={'edgecolor': 'black'})

plt.title('Popularity Top 5 Languages')
plt.tight_layout()

plt.savefig('PieChartPopLang.png')
plt.show()
```



In [76]:

```
labels = ['JavaScript', 'HTML', 'SQL', 'Python', 'Java', 'C++']
slices = [LearnJavaScript, learnHTML, learnSQL, learnPy, learnJava, learnCpp]
explode = [0, 0, 0, 0.1, 0, 0]

plt.pie(slices, labels=labels, explode=explode, shadow=True, startangle=120, autopct=
       wedgeprops={'edgecolor': 'black'})

plt.title('Planning to Learn')
plt.tight_layout()
plt.savefig('PieChartLearnLang.png')
plt.show()
```



Conclusions Drawn:

▪ The United States has a noticeably higher mean and median salary. The UK, Canada and Germany all have similar mean and median salaries. India claimed the fifth highest median salary, though the median salary drops off considerably. (India's median salary is 9% of the US median salary and just under 20% of the world median salary)

▪ The median age of the respondent is 30 years old. This aligns with expectations

▪ A higher percentage of younger respondents 20-35 years old tend to already know Python. Over the age of 35, the percentage of respondents that already know Python drop off

▪ Based on responses to this survey, Python is a popular upcoming programing language. This is not surprising given the ease and versatility of the Python programming language. The increase in desire to learn this language in 2021, compared to the relatively low percentage of respondents that currently know Python help support this conclusion.

End Report 3/18/21 completed by Ryan Olsen