Open in Colab

After clicking the "Open in Colab" link, copy the notebook to your own Google Drive before getting started, or it will not save your work

BYU CS 180 Lab 4: Data Wrangling

Introduction:

In this lab, we will analyze college football data from the years 2016-2020. The data is spread across multiple files and will require a bit of “data wrangling”.

Once we have cleaned and processed the data, we will put our statistics knowledge to good use by digging a little deeper than simple summary statistics.

Very rarely are data scientists handed a pristine data set ready for analysis. More often than not, quite a bit of work is required to clean and preprocess the data so that it's ready for analysis.

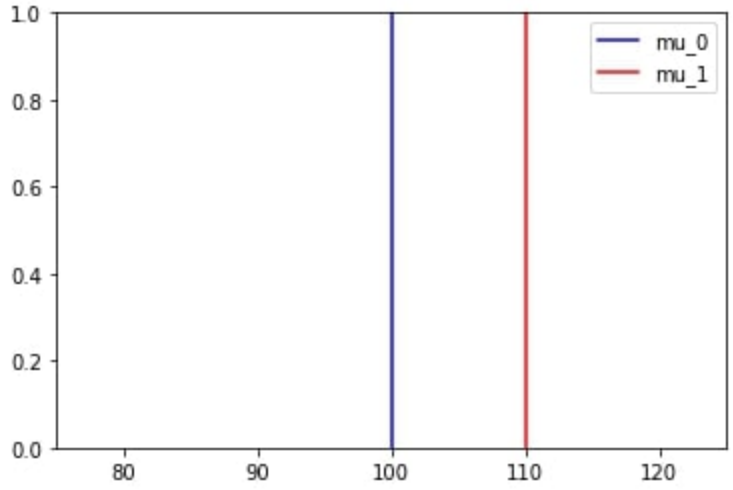
Data wrangling, also called **data cleaning**, **data remediation**, or **data munging**, refers to a variety of processes designed to transform raw data into more readily usable formats. The exact methods differ from project to project depending on the data you’re leveraging and the goal you’re trying to achieve.

Some examples of data wrangling include:

- Merging multiple data sources into a single dataset for analysis.
- Identifying gaps in data (for example, empty cells in a spreadsheet) and either filling or deleting them.
- Deleting data that's either unnecessary or irrelevant to the project you're working on.
- Identifying extreme outliers in data and either explaining the discrepancies or removing them.

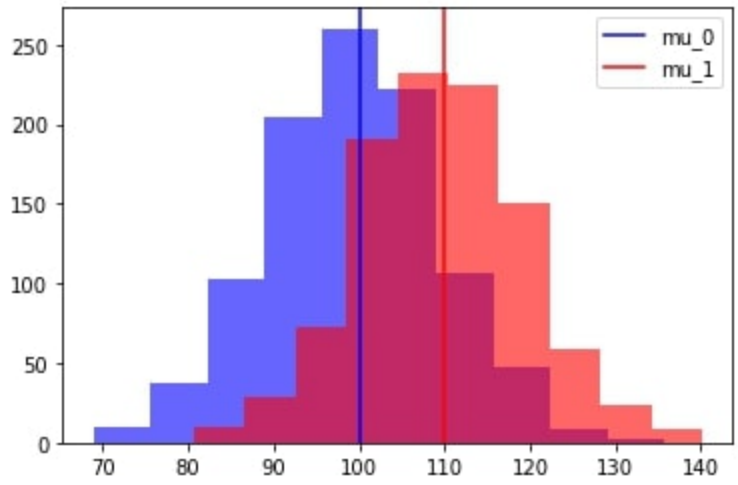
Measures of Variability

We are very used to comparing point estimates. For example, in order to tell if one thing is better than the other, we may look at the average of each over time. Consider the following plot comparing the average of two groups, mu_0 and mu_1.



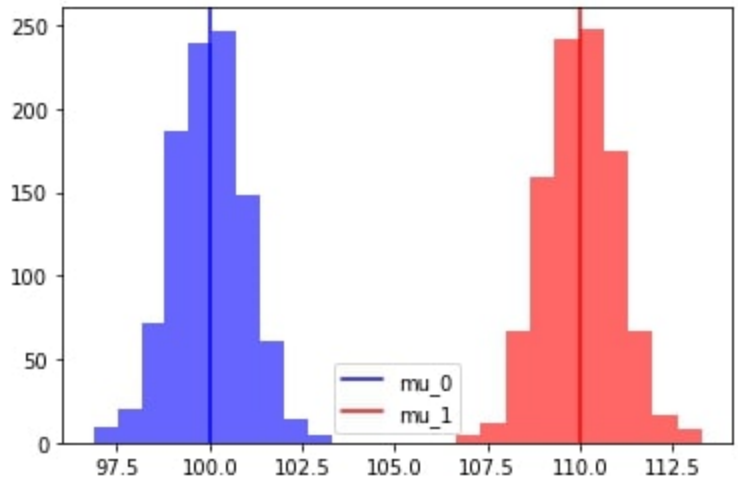
In this case, we might conclude that mu_1 is greater because it's average is higher. However, we are ignoring an important aspect of the data: its variability.

When we plot the **variability and the mean**, we observe the following plot:



Now the distributions of mu_1 and mu_0 are so close that it's difficult to say with certainty that mu_1 is better. For any random draw of mu_1, it's roughly a 50-50 chance to actually be greater than a random draw of mu_0.

Now suppose we observe the following plot:



While the means are the same as before, the variability of the two distributions are now very different. We can say with a high degree of certainty that mu_1 is higher.

Keep this idea in mind as you analyze the dataset for this lab.

Data:

Data from the 2016, 2017, 2018, 2019, and 2020 college football seasons are also available on the course GitHub:

In [1]:
```
# Download the data located at the following URLs
cfb16_url = "https://raw.githubusercontent.com/rhodes-byu/cs180-winter25/refs/heads
cfb17_url = "https://raw.githubusercontent.com/rhodes-byu/cs180-winter25/refs/heads
cfb18_url = "https://raw.githubusercontent.com/rhodes-byu/cs180-winter25/refs/heads
```

```
cfb19_url = "https://raw.githubusercontent.com/rhodes-byu/cs180-winter25/refs/heads
cfb20_url = "https://raw.githubusercontent.com/rhodes-byu/cs180-winter25/refs/heads
```

Exercise 1: Yearly Counts

Exercise Question

Read in the files, and add a year column to each file (from the original .csv file name).

In [2]:
```
# Write your code to read in the files and add the year values from each csv:
import pandas as pd
cfb16 = pd.read_csv(cfb16_url)
cfb16['Year'] = 2016
cfb17 = pd.read_csv(cfb17_url)
cfb17['Year'] = 2017
cfb18 = pd.read_csv(cfb18_url)
cfb18['Year'] = 2018
cfb19 = pd.read_csv(cfb19_url)
cfb19['Year'] = 2019
cfb20 = pd.read_csv(cfb20_url)
cfb20['Year'] = 2020
```

Exercise 2: Data Aggregation

Exercise Question

Combine every file into a single dataframe.

In [3]:
```
# Write your code to combine all of the csvs into one dataframe here:
combined_cfb = pd.concat([cfb16, cfb17, cfb18, cfb19, cfb20], ignore_index=True)
```

Exercise 3: Conference Search

Exercise Question

Create a `conference` field by parsing the `team` column.

Example:

Team	Conference
Penn State University (Big Ten)	Big Ten

In [4]:
```
# Write your code to parse the conference from the team name:
combined_cfb['Conference'] = combined_cfb['Team'].str.extract(r'\((.*?)\)')
```

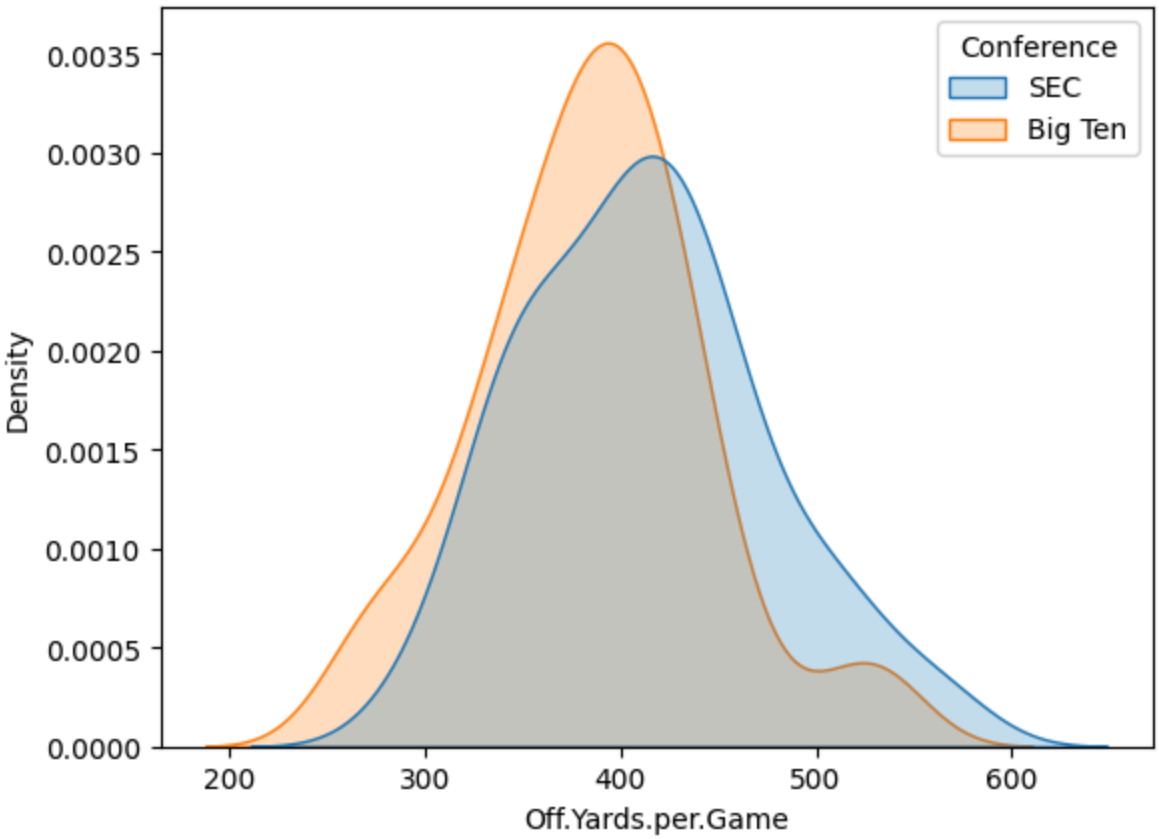
Exercise 4: Big Ten Vs. South Eastern

Exercise Question 4a: Offense

- Is there a statistical difference between the Big Ten Conference and the South Eastern Conference in terms of `Off.Yards.per.Game`? Use a seaborn KDE plot to create a figure. Comment on the difference in means and the overlap of distributions.
- Do the same as above for `Off.TDs`.

In [ ]:
```
# Write the code for the statistical differences for Off.Yards.per.Game:
import seaborn as sns
import matplotlib.pyplot as plt


compared_cfb = combined_cfb[combined_cfb['Conference'].isin(['Big Ten', 'SEC'])]
sns.kdeplot(data=compared_cfb, x='Off.Yards.per.Game', hue='Conference', fill=True,
plt.show()
```

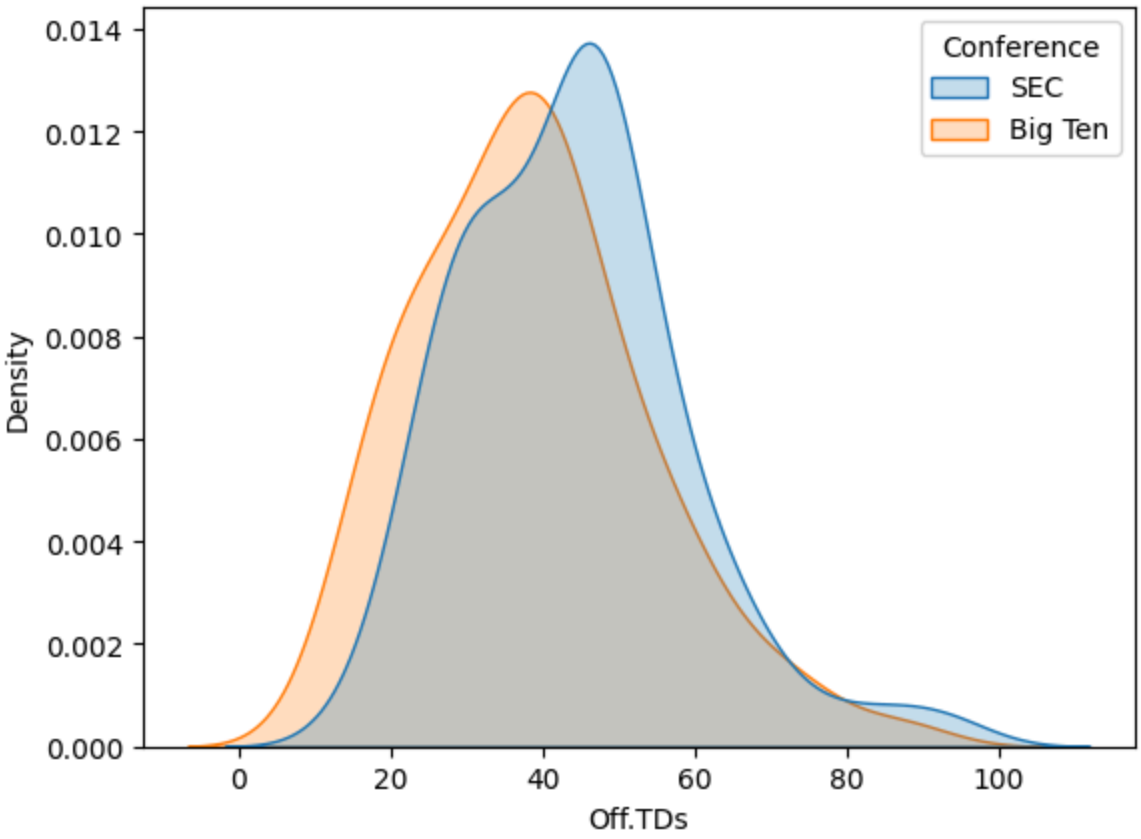


(Comment on the difference in means and the overlap of distributions here)

I think the overlap is high in that both conferences perform similarly in a lot of their games. But the density of the SEC is more spread out and further to the left meaning that they have

a higher average Off.Yards.per.Game (I would say that both teams have the same spread width).

In [ ]:
```
# Write the code for the statistical differences for Off.TDs:
sns.kdeplot(data=compared_cfb, x='Off.TDs', hue='Conference', fill=True, warn_singu
plt.show()
```



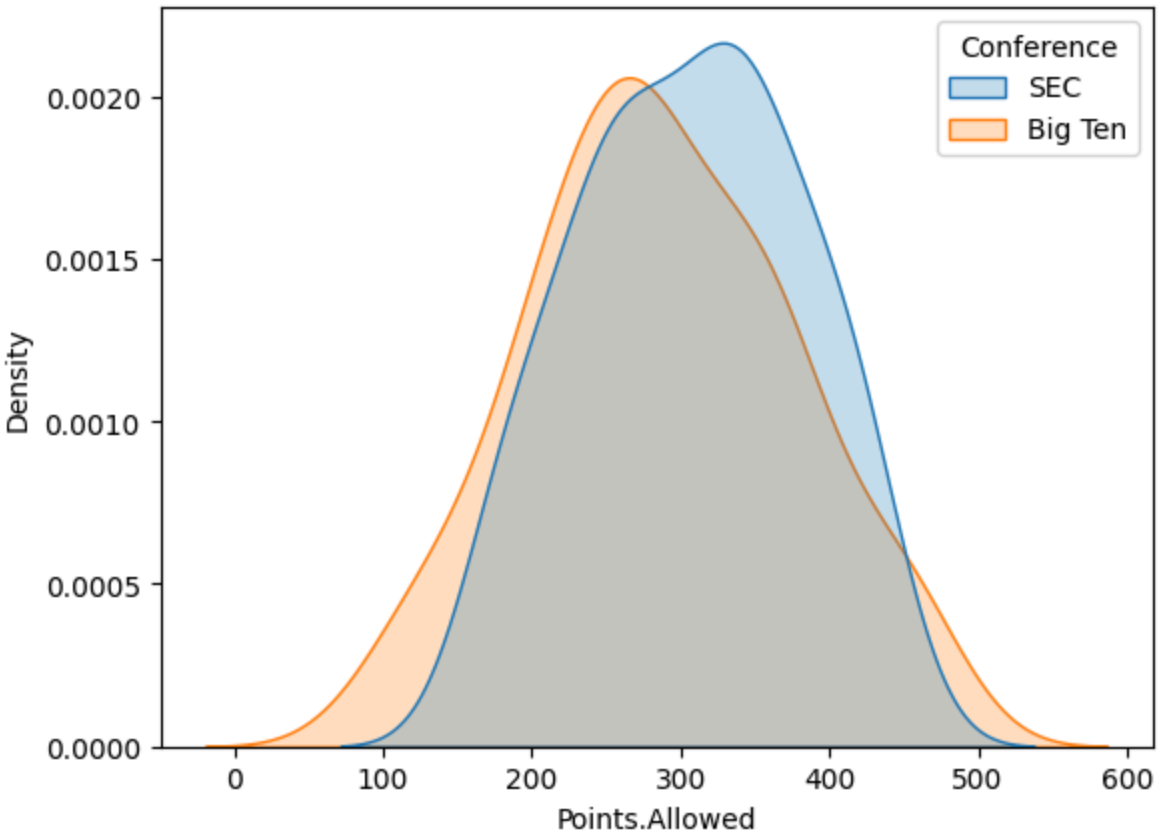
(Comment on the difference in means and the overlap of distributions here)

Similar to Off.Yards.per.Game, both conferences have the same spread and general shape, but the SEC is further to the right meaning that they have a higher average of offensive touchdowns per game.

Exercise Question 4b: Defense

- Is there a statistical difference between the Big Ten Conference and the South Eastern Conference in terms of `Points.Allowed`? Use a seaborn KDE plot to create a figure. Comment on the difference in means and the overlap of distributions.
- Do the same as above for `Opp.Pass.Yds.Allowed` and `Opp.Rush.Yards.Allowed`.

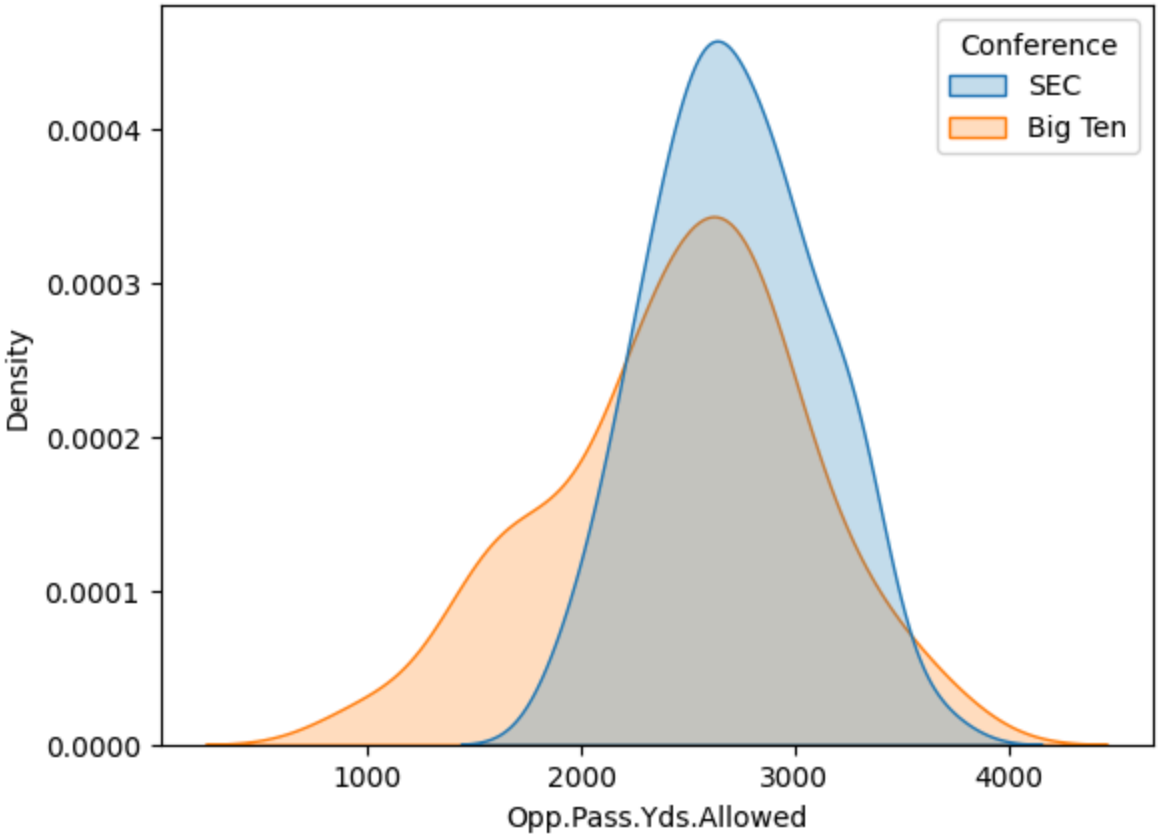
In [7]:
```
# Write the code for the statistical differences for Points.Allowed:
sns.kdeplot(data=compared_cfb, x='Points.Allowed', hue='Conference', fill=True, war
plt.show()
```



(Comment on the difference in means and the overlap of distributions here)

It looks like both leagues are continuing to be relatively even in their football skills, but since this is points allowed the fact that the SEC is further to the left is actually a detriment and I would guess that the Big Ten did better overall in this statistic.

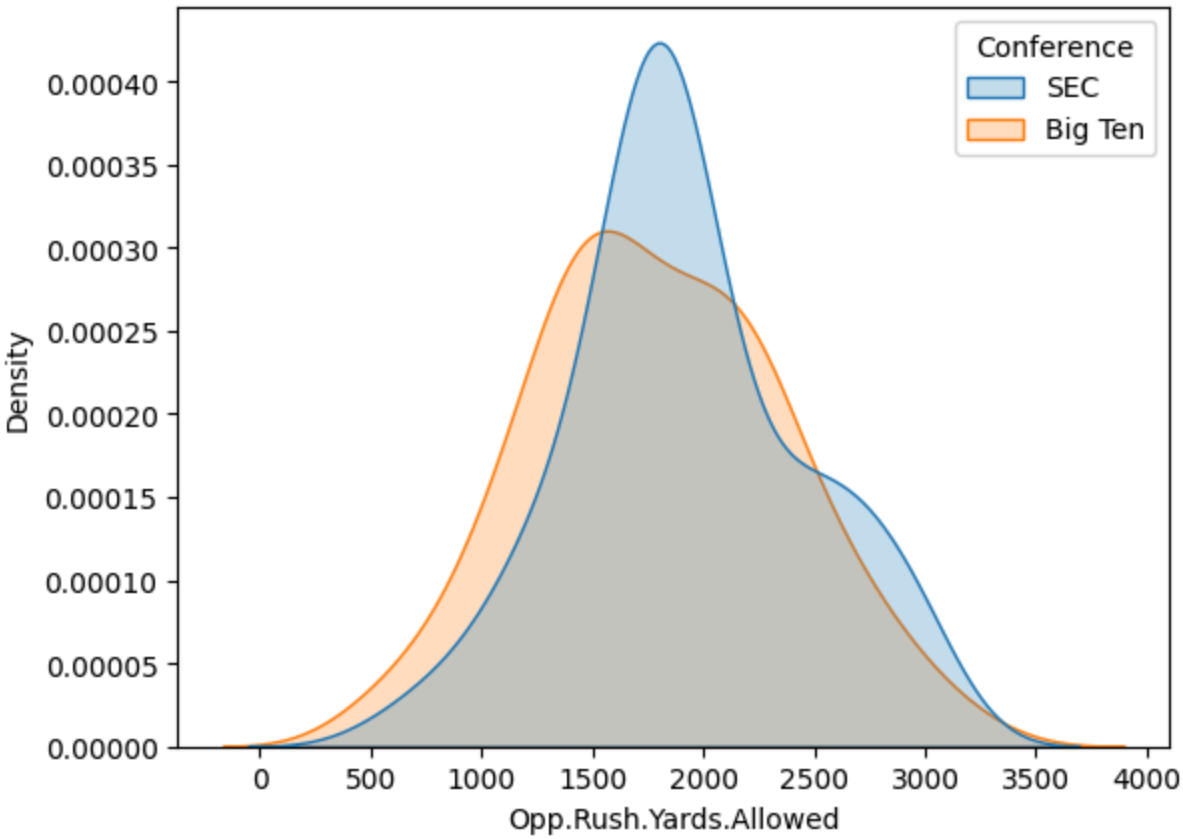
In [8]:
```
# Write the code for the statistical differences for Opp.Pass.Yds.Allowed:
sns.kdeplot(data=compared_cfb, x='Opp.Pass.Yds.Allowed', hue='Conference', fill=Tru
plt.show()
```



(Comment on the difference in means and the overlap of distributions here)

Unlike the last few graphs this graph is a lot more separated between the conferences, and since less is better in Opp.Pass.Yrds.Allowed, I would confidently say that the Big Ten did better overall that the SEC.

In [9]:
```
# Write the code for the statistical differences for Opp.Rush.Yards.Allowed:
sns.kdeplot(data=compared_cfb, x='Opp.Rush.Yards.Allowed', hue='Conference', fill=T
plt.show()
```



(Comment on the difference in means and the overlap of distributions here)

Similarly to the last graph, less is better with this one and while the SEC and Big Ten have the same spread, I would say that because the SEC has a higher peak they did worse in this category.

Exercise 5: Offense Statistics

Exercise Question 5a: Offense

Is the offense changing over time? Create some plots showing the average offensive production over time (each year). Include an estimate of the variability in your figures.

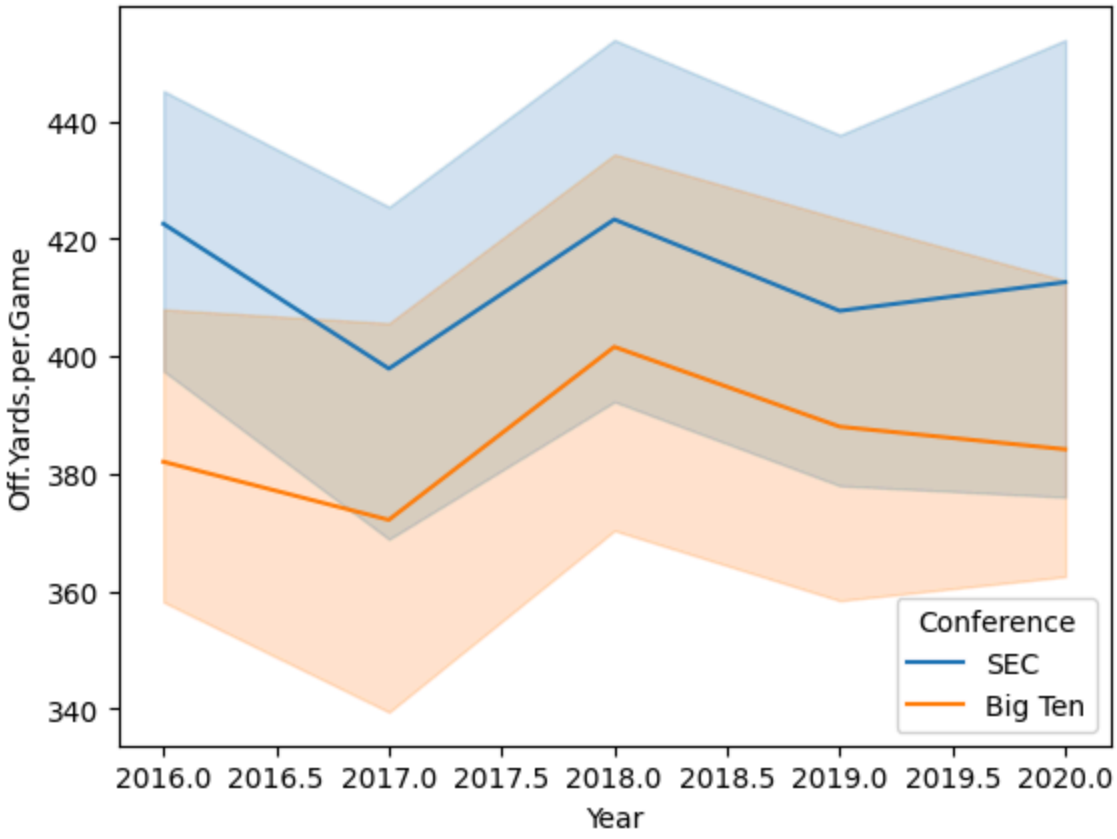
Comment on whether any trends you see are likely to be true or spurious.

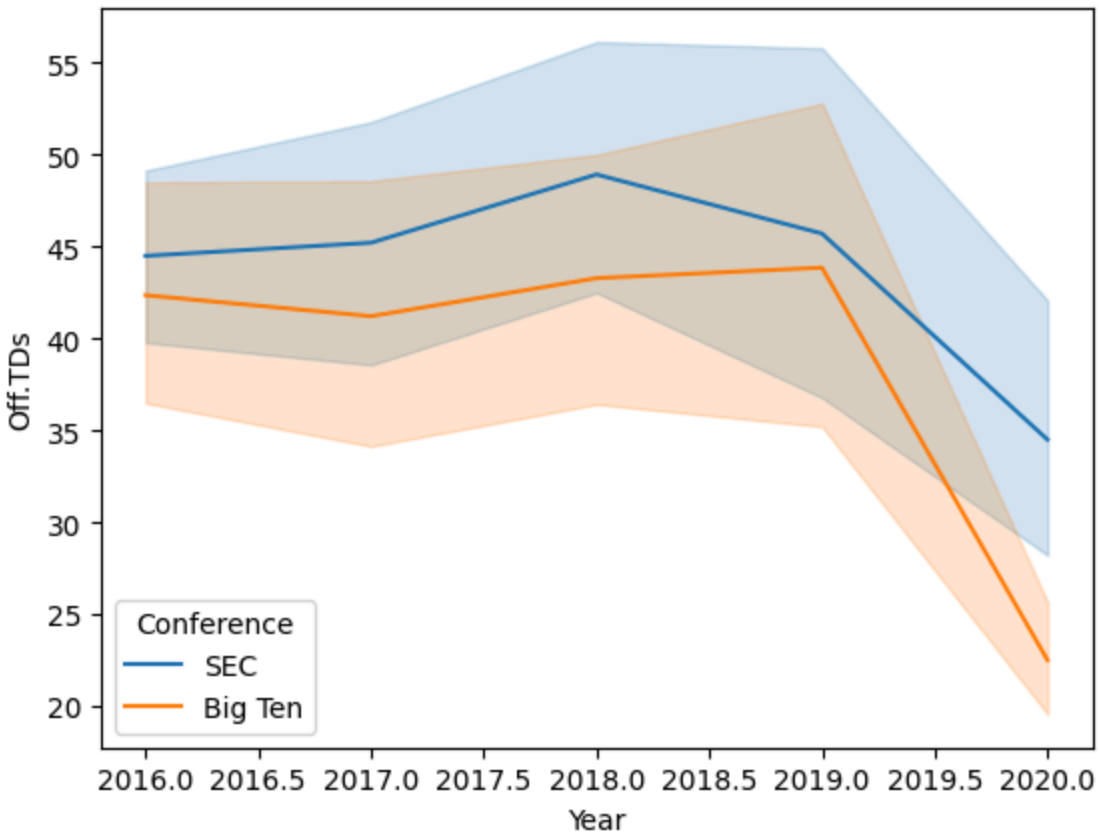
Create a plot for the following metrics:

- `Off.Yards.per.Game`
- `Off.TDs`

In [10]:
```
# Create a few plots showing how each metric changed over time:
sns.lineplot(data=compared_cfb, x='Year', y='Off.Yards.per.Game', hue='Conference')
plt.show()
```

```
sns.lineplot(data=compared_cfb, x='Year', y='Off.TDs', hue='Conference')
plt.show()
```





(Comment on any trends you see here)

Off.Yards.per.Game: Both teams have a drop in 2017 and had a great year in 2018 but overall tended to stay stead throughout the years.

Off.TDs: It looks like both conferences were holding pretty steady (SEC had a good 2018) but starting in 2019 both conferences really started dropping in stats.

I don't understand the instruction about whether I think any trends are true or spurious. I am working under the assumption that the data was collected correctly, therefore I assume the trends to all be true. Now whether there is some real world cause for the trends, I would guess the drop for Off.TDs was caused by COVID starting up.

Exercise Question 5b: Defense

Is the defense changing over time? Create some plots showing the average defensive production over time (each year). Include an estimate of the variability in your figures.

Comment on whether any trends you see are likely to be true or spurious.

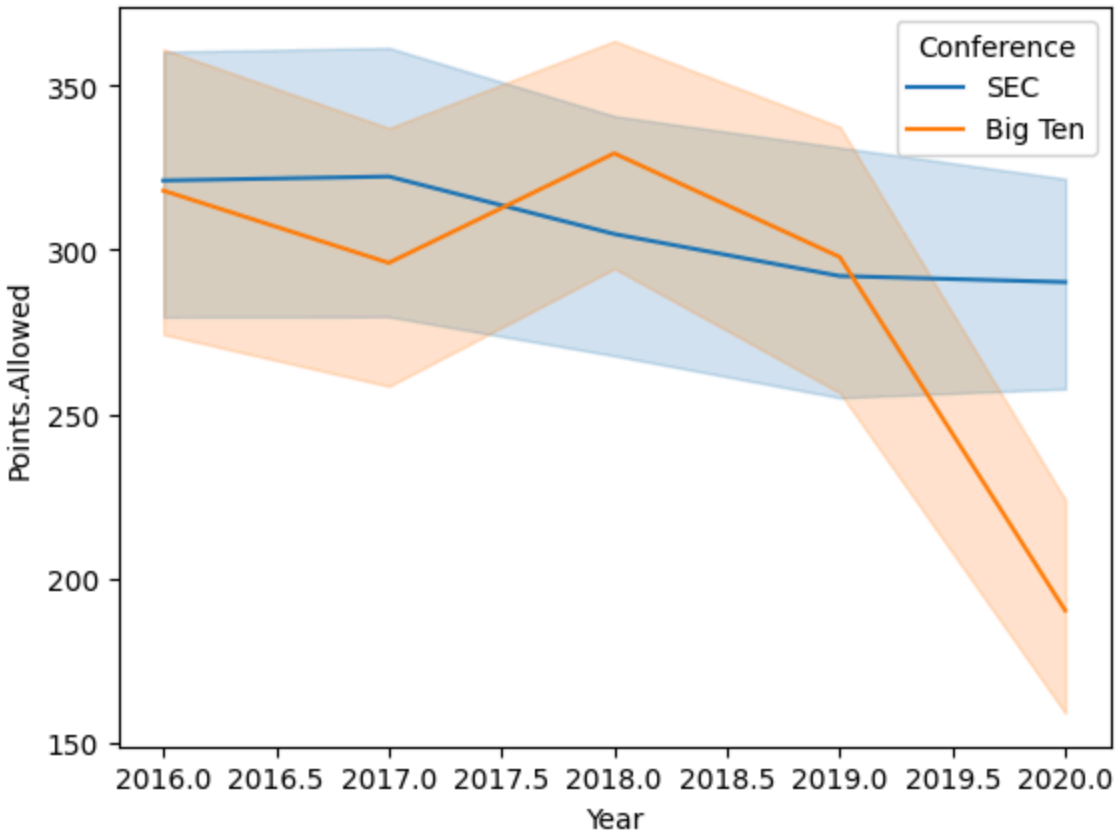
Create a plot for the following metrics:

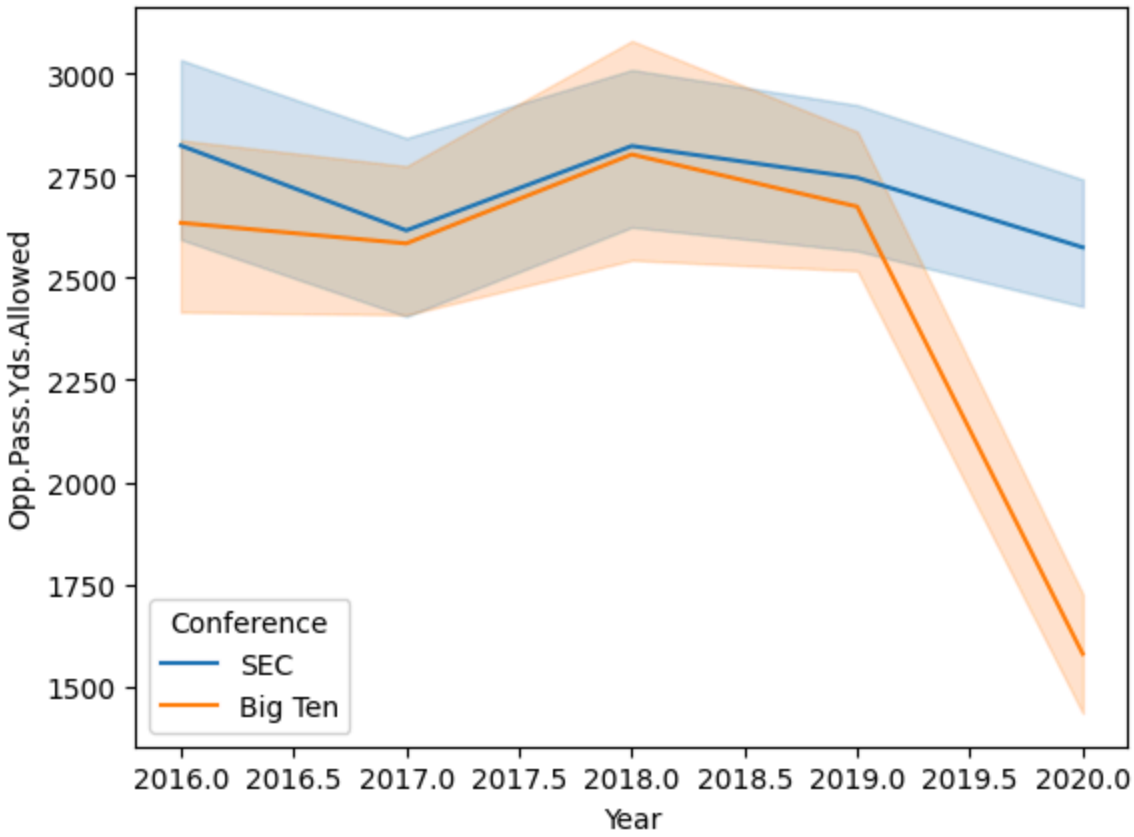
- `Points.Allowed`
- `Opp.Pass.Yds.Allowed`
- `Opp.Rush.Yards.Allowed`

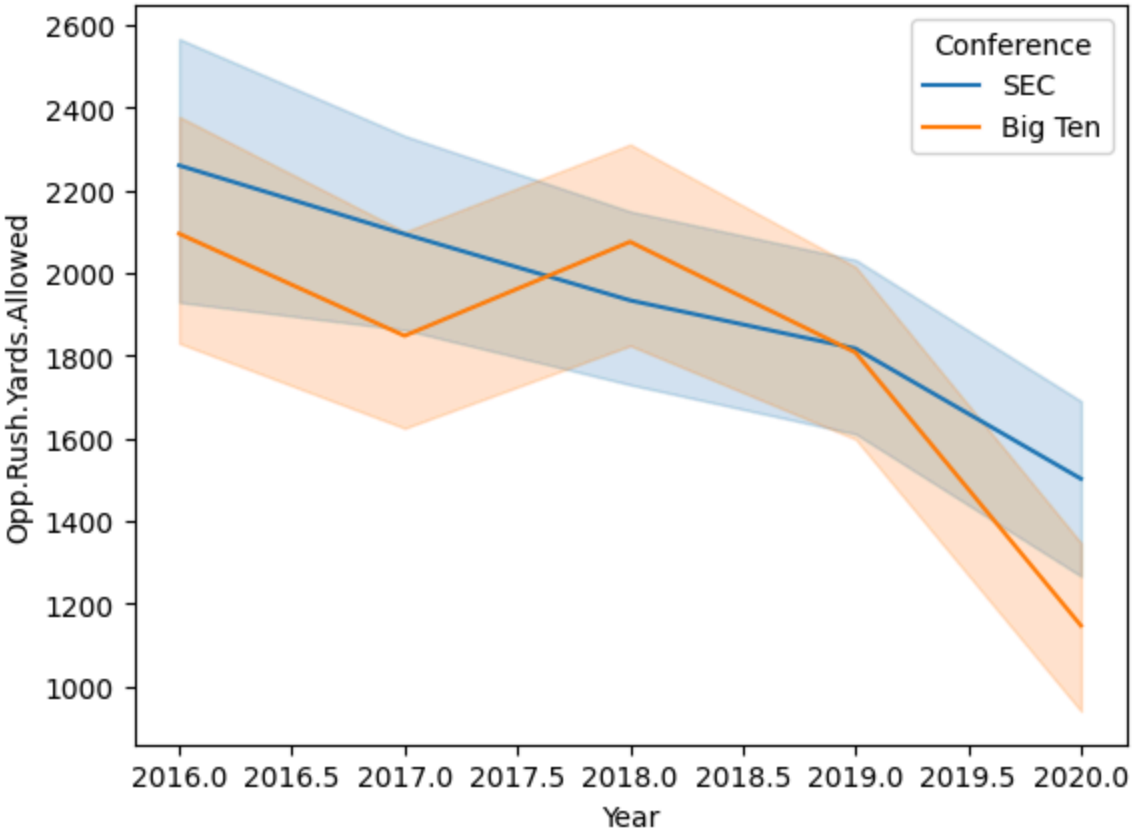
In [11]:
```
# Create a few plots showing how each metric changed over time:
sns.lineplot(data=compared_cfb, x='Year', y='Points.Allowed', hue='Conference')
plt.show()

sns.lineplot(data=compared_cfb, x='Year', y='Opp.Pass.Yds.Allowed', hue='Conference
plt.show()

sns.lineplot(data=compared_cfb, x='Year', y='Opp.Rush.Yards.Allowed', hue='Conferen
plt.show()
```


Conference
SEC
Big Ten
350
300
250
200
150
Points.Allowed
2016.0 2016.5 2017.0 2017.5 2018.0 2018.5 2019.0 2019.5 2020.0
Year


3000
2750
2500
2250
2000
1750
1500
Opp.Pass.Yds.Allowed
Conference
SEC
Big Ten
2016.0 2016.5 2017.0 2017.5 2018.0 2018.5 2019.0 2019.5 2020.0
Year



(Comment on any trends you see here)

All 3 plots are similar enough that I will just use one comment. At first the defence of the SEC and the Big Ten flowed in reverse to one another, one goes up the other goes down. But around 2018 both conference's stats start to go downward, but in the Big Ten's case they plummeted in points allowed (which means they were doing way better than the SEC).