How Data Science Assists Sports: Jupyter Notebook, Analytics, and the 3-point Revolution

Chris Rawles, Senior Data Scientist @ Pivotal

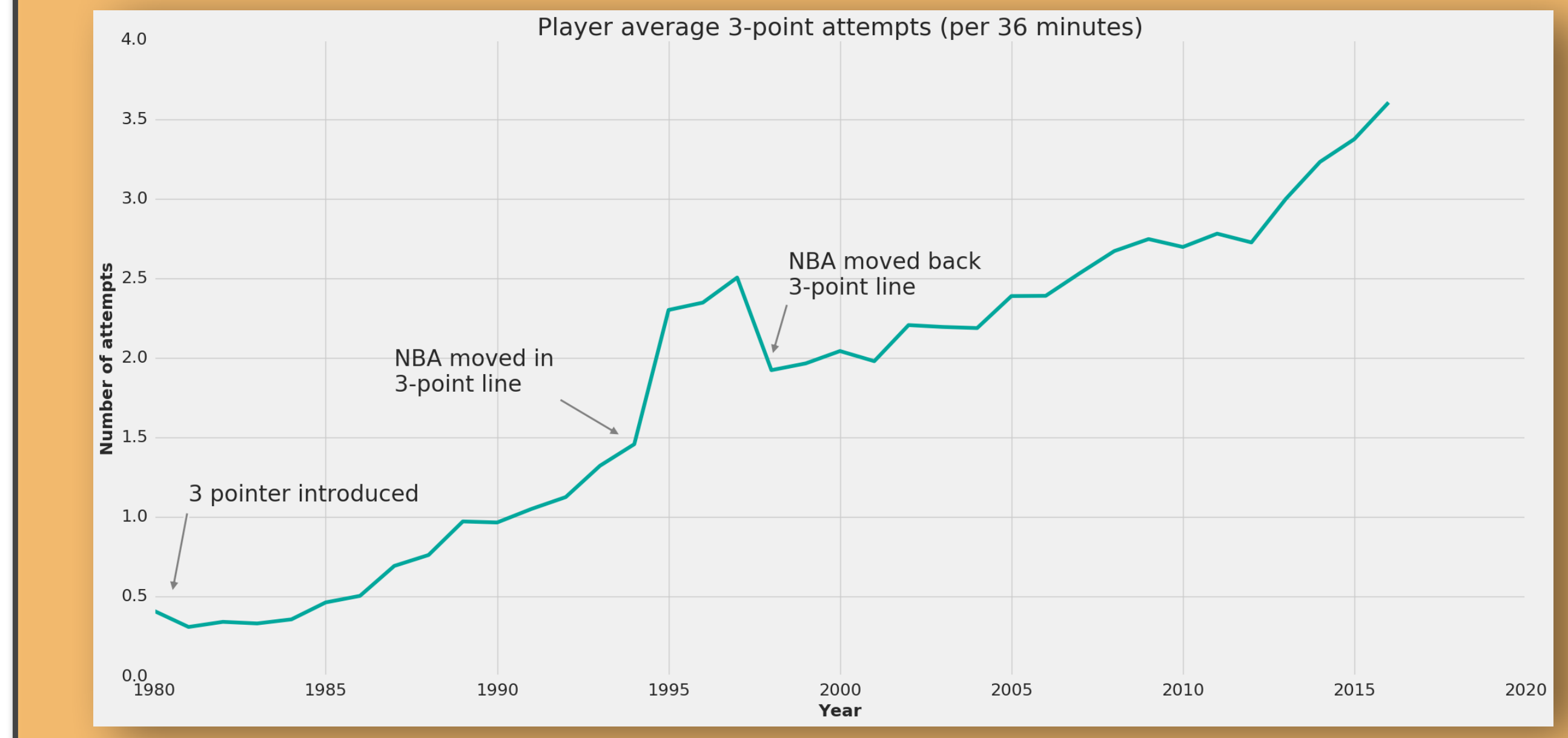
Pivotal™

Summary

- Analytics have become a major tool in both amateur and professional basketball, shaping how the sport is played
- Both the NBA and the NCAA has skewed towards taking more 3-point shots due to their high efficiency as measured by points per field goal attempt
- Apache Spark, Python, and Jupyter Notebook for analysis

Three-point shooting is on the rise in the NBA

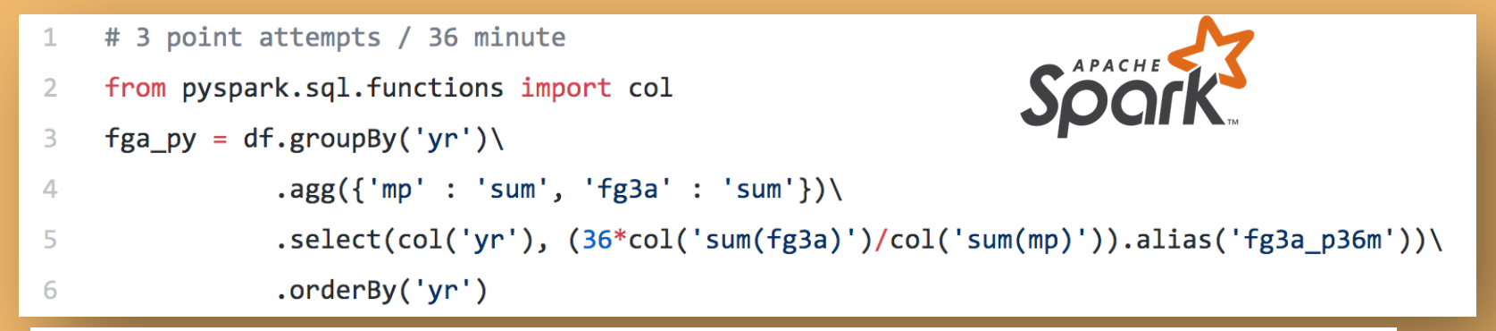
The Evolution of 3-point shooting since 1979



We can see a steady rise in the number of 3 point attempts since the shot's introduction in the 1979-80 season, along with a blip in number of attempts during the period in the mid 90's when the NBA moved the line in a few feet.

Apache Spark

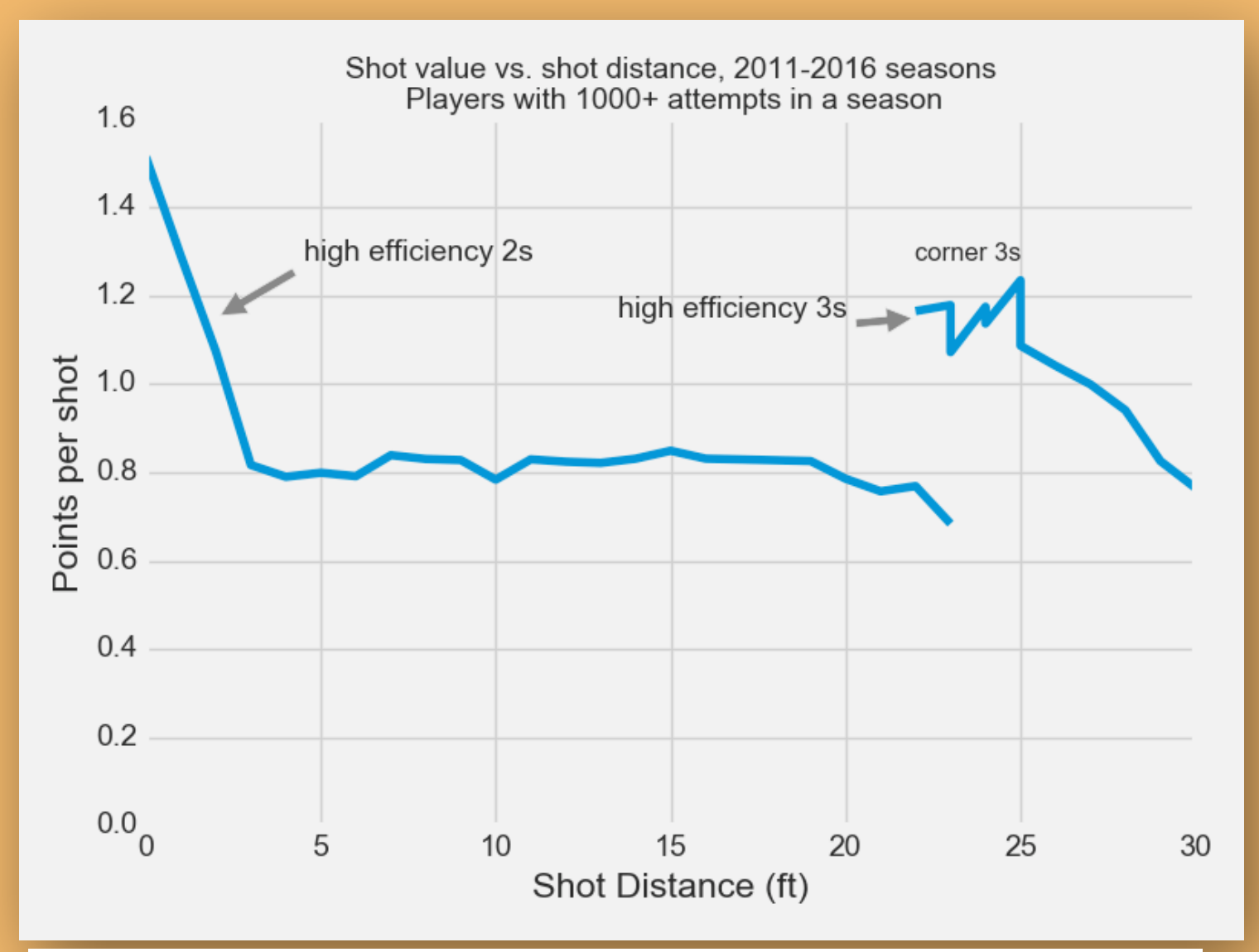
- Spark is a distributed computing engine for processing large amounts of data
- The Spark DataFrame API has Pandas-like syntax
- The Spark Python API, PySpark, interfaces quite nicely with Jupyter Notebook



```
# 3 point attempts / 36 minute
from pyspark.sql.functions import col
fga_py = df.groupBy('yr')\
           .agg({'mp' : 'sum', 'fg3a' : 'sum'})\
           .select(col('yr'), (36*col('sum(fg3a)')/col('sum(mp)')).alias('fg3a_p36m'))\
           .orderBy('yr')
```

Example PySpark code for computing 3-point shooting per 36 minutes

Why the 3-pointer is the most efficient shot in basketball



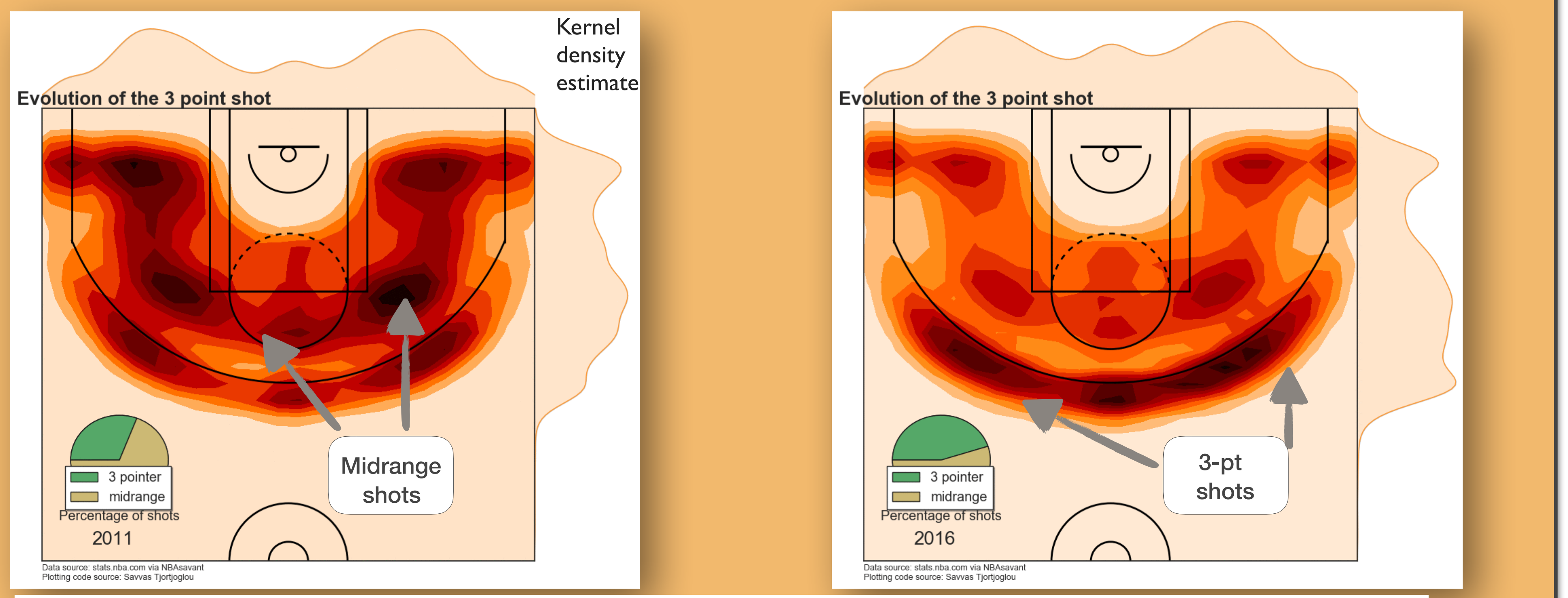
Close 3-point attempts are among the most efficient shots in the league, on par with shots taken close to the basket. It's no wonder that accurate 3-point shooting is among the most coveted talents in the NBA today!

Midrange shots are being reduced as 3-point shots are increasing

- Midrange shots, shots too far away to be layups yet too close to be 3-point shots, have a low expected value
- As basketball has evolved, teams are shooting less midrange shots and more 3-point shots
- This trend is evident even in the past 5 years and is still continuing

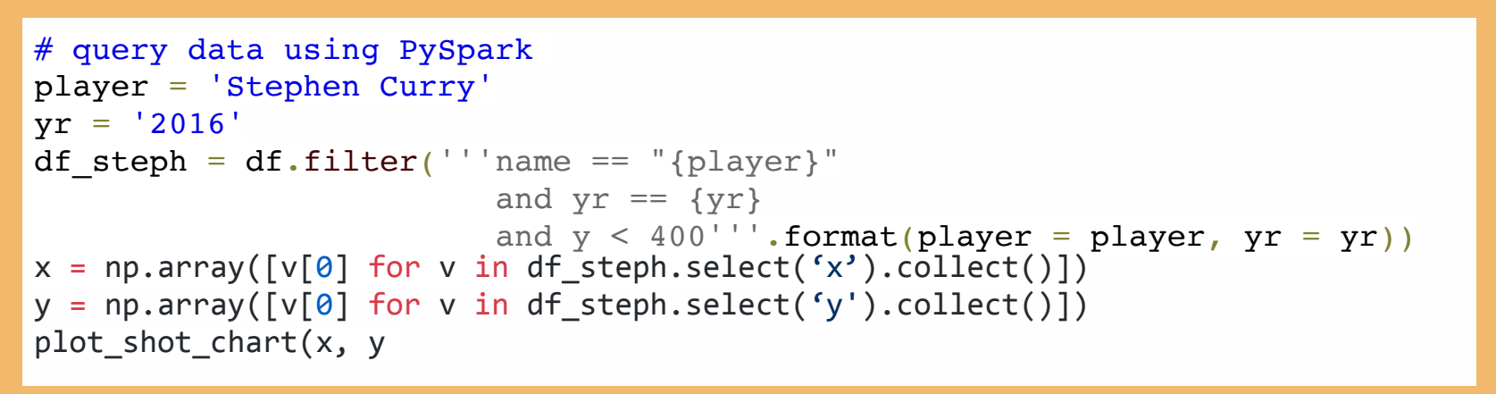
2010-2011

2015-2016

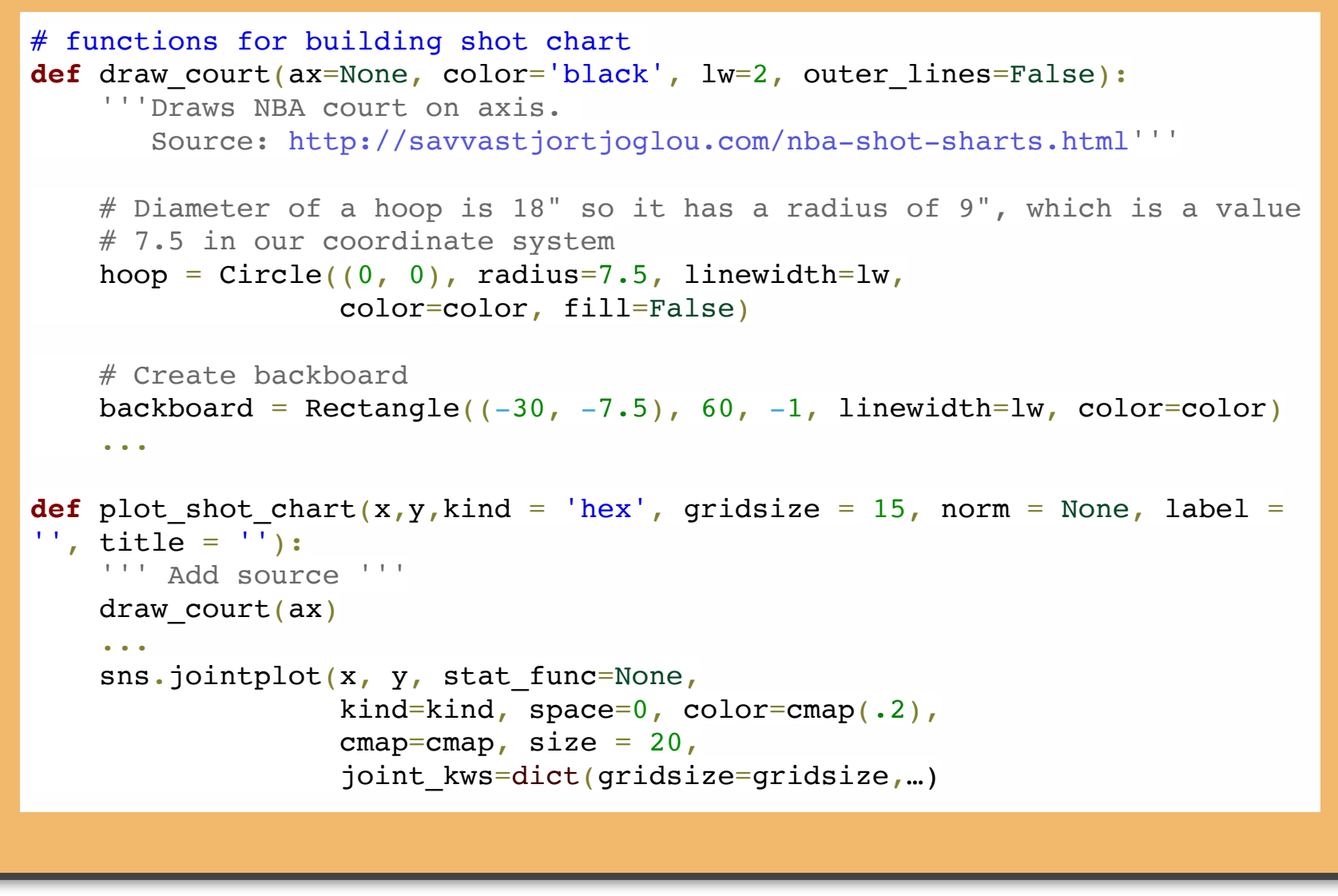


Shot density charts for the (left) 2010-2011 season and (right) 2015-2016 season. Darker colors indicate a higher number of shots taken from that area. The 2010-2011 season has a much higher shot density in the midrange area inside the arc than the 2015-2016 season.

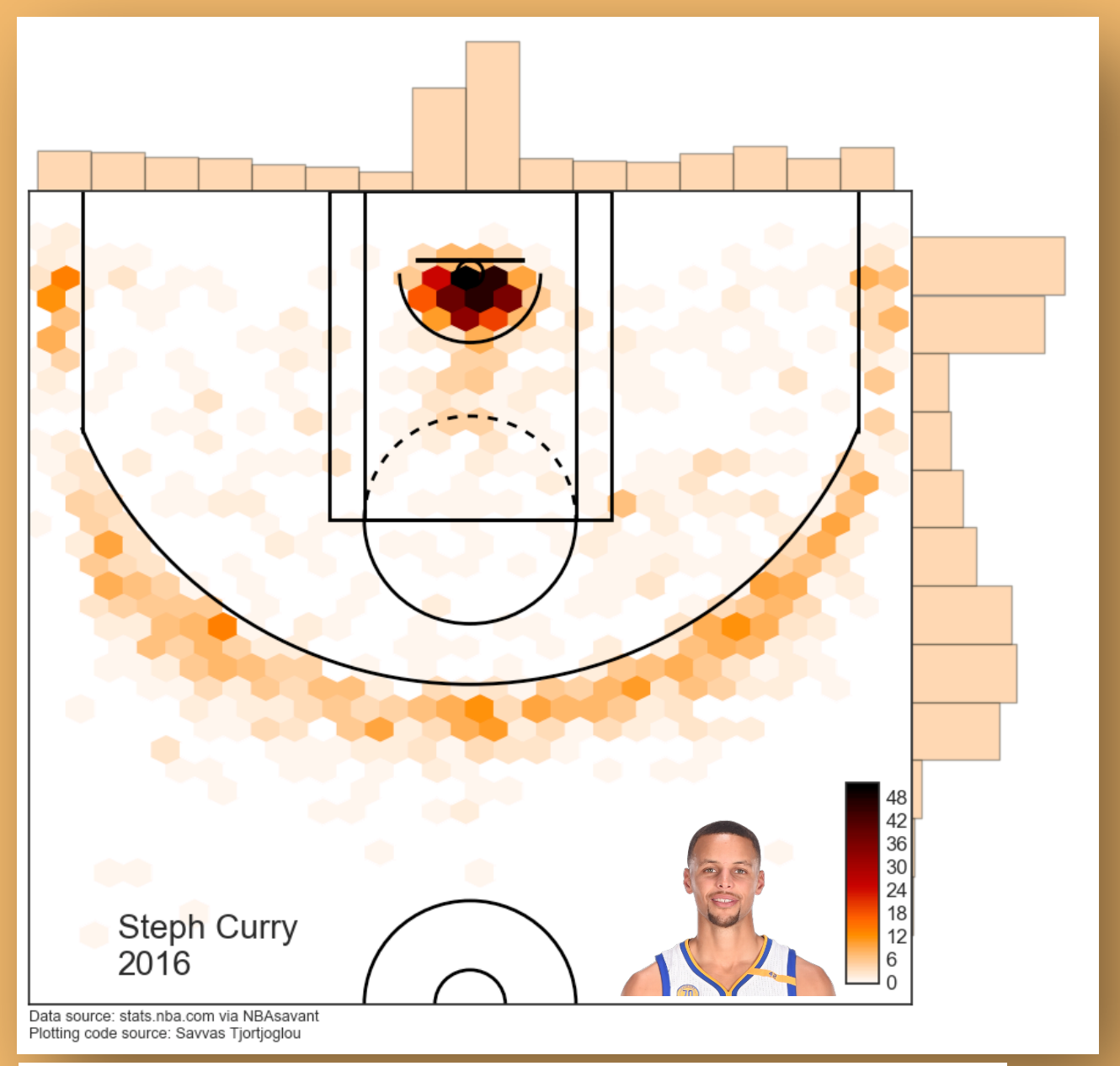
Building a shot chart using Spark, Seaborn, and Matplotlib



```
# query data using PySpark
player = 'Stephen Curry'
yr = '2016'
df_steph = df.filter('''name == "{player}"
                        and yr == {yr}
                        and y < 400'''.format(player = player, yr = yr))
x = np.array([v[0] for v in df_steph.select(‘x’).collect()])
y = np.array([v[0] for v in df_steph.select(‘y').collect()])
plot_shot_chart(x, y
```



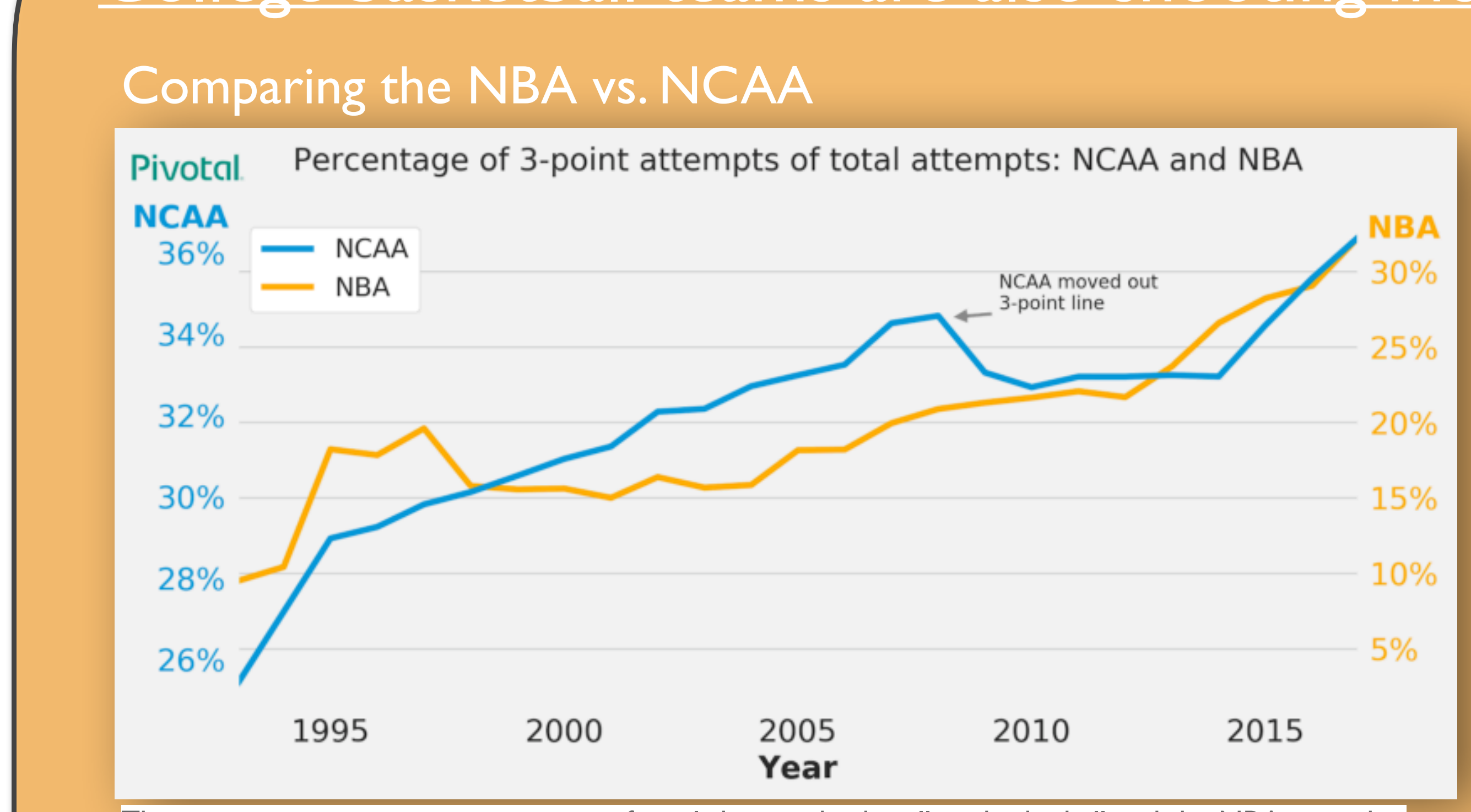
```
# functions for building shot chart
def draw_court(ax=None, color='black', lw=2, outer_lines=False):
    '''Draws NBA court on axis.
       Source: http://savvastjortjoglou.com/nba-shot-sharts.html'''

    # Diameter of a hoop is 18" so it has a radius of 9", which is a value
    # 7.5 in our coordinate system
    hoop = Circle((0, 0), radius=7.5, linewidth=lw,
                  color=color, fill=False)

    # Create backboard
    backboard = Rectangle((-30, -7.5), 60, -1, linewidth=lw, color=color)
    ...

def plot_shot_chart(x,y,kind = 'hex', gridsize = 15, norm = None, label =
'', title = ''):
    ''' Add source '''
    draw_court(ax)
    ...
    sns.jointplot(x, y, stat_func=None,
                  kind=kind, space=0, color=cmap(.2),
                  cmap=cmap, size = 20,
                  joint_kws=dict(gridsize=gridsize,…)
```



Steph Curry's historic 3-point shooting 2015-2016 season, by number of shots at each location on the court.

College basketball teams are also shooting more 3-pointers

Comparing the NBA vs. NCAA

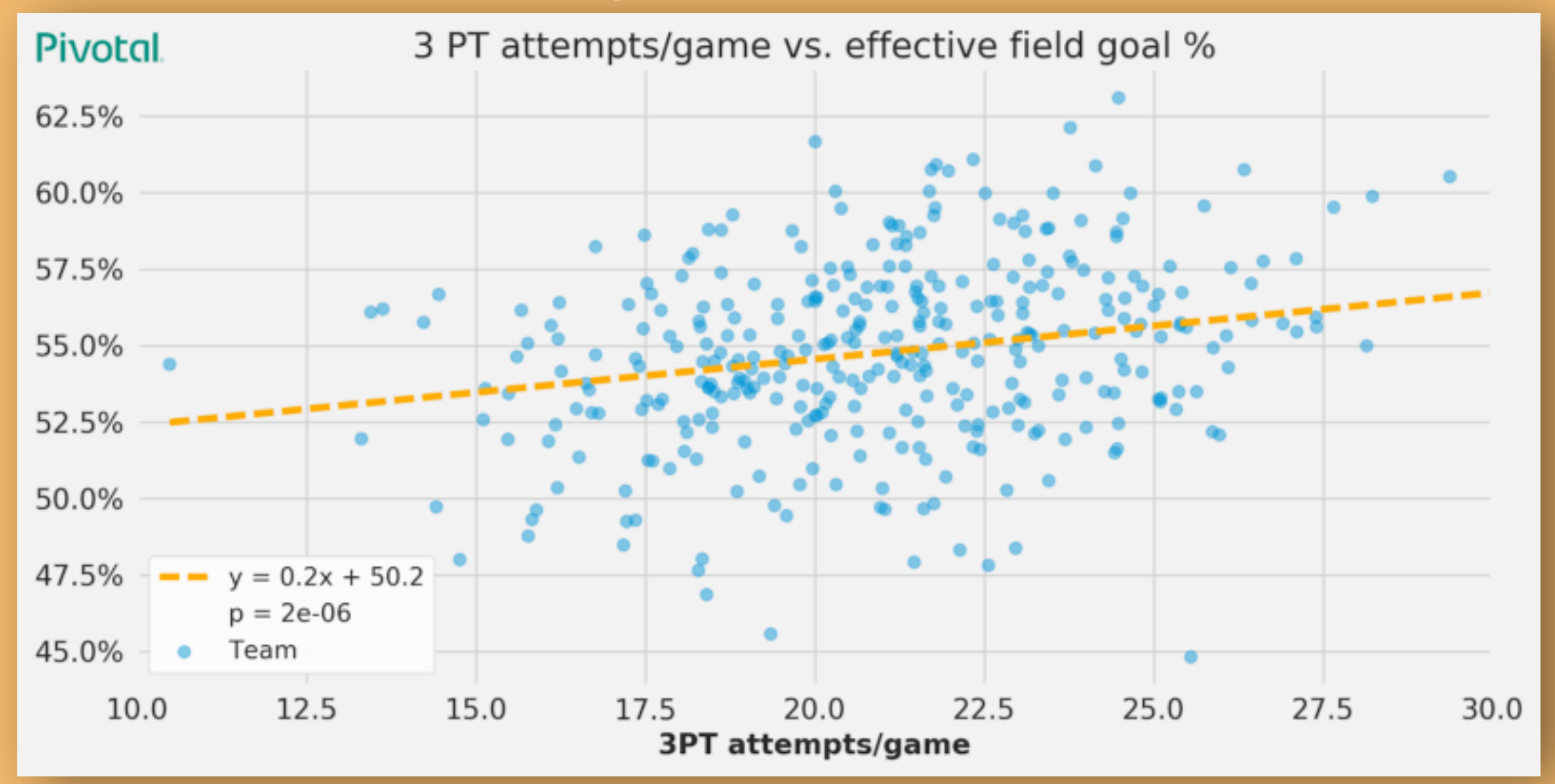


Three point attempts, as a percentage of total shots, in both college basketball and the NBA over the past 25 years

- College teams shot more 3-point shots than ever in the 2016–17 season with a record-breaking 36 percent of total shots coming from downtown.
- Three point attempts have been on the rise nearly every year in college basketball, with a notable exception of the 2008 season when the NCAA moved the college line out one foot to 20 feet 9 inches.

- Effective shooting percentage (eFG%) provides a more complete picture of a team's or player's shooting.
- eFG% adjusts for the fact that while a 3-point shoot typically returns a lower field goal percentage (FG%), it should be valued higher because it's worth 50 percent more than a 2-point shot.
- Shooting a lot of 3-pointers isn't the only way to achieve a high eFG% (high 3-point accuracy and close-to-the-basket 2-pointers help a lot), although the two are certainly correlated.

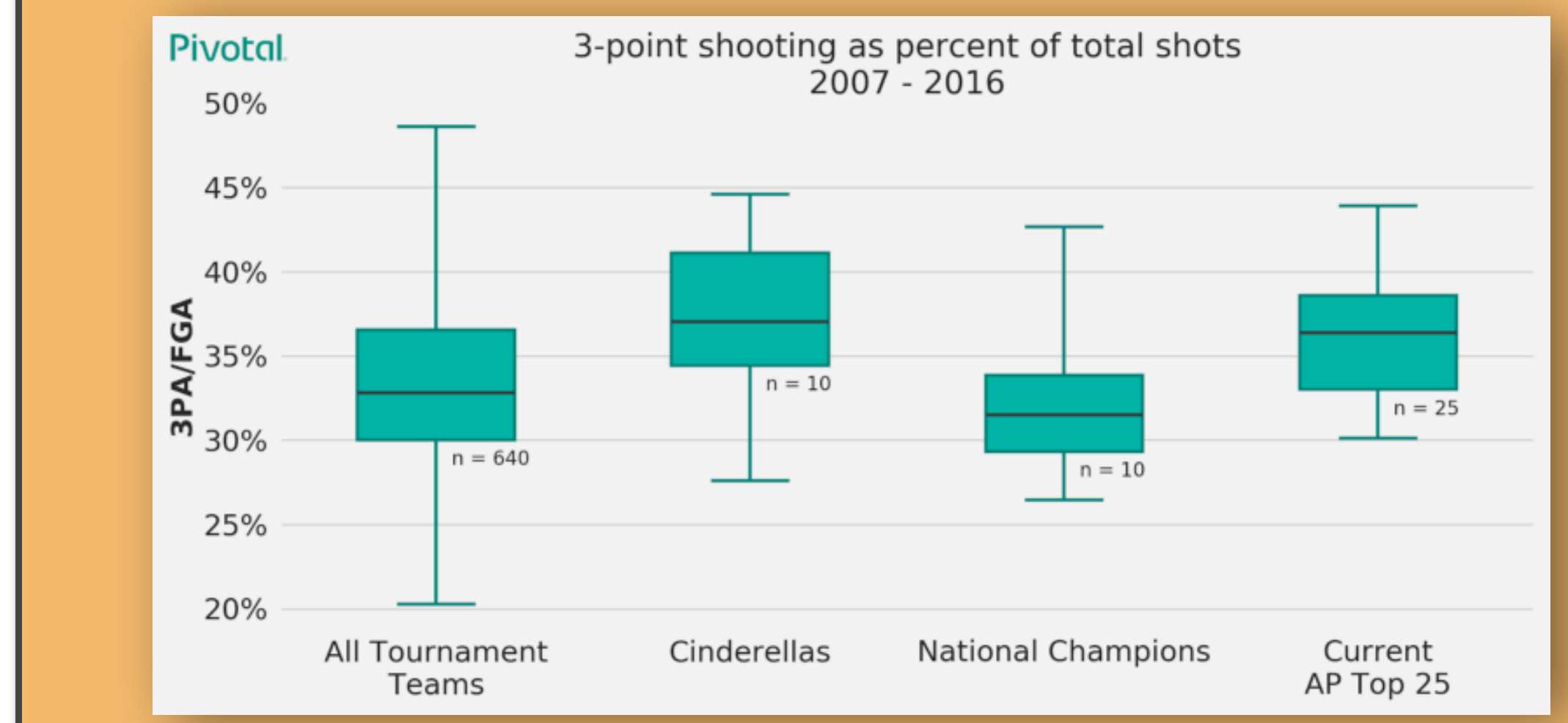
2016-2017 NCAA Regular Season



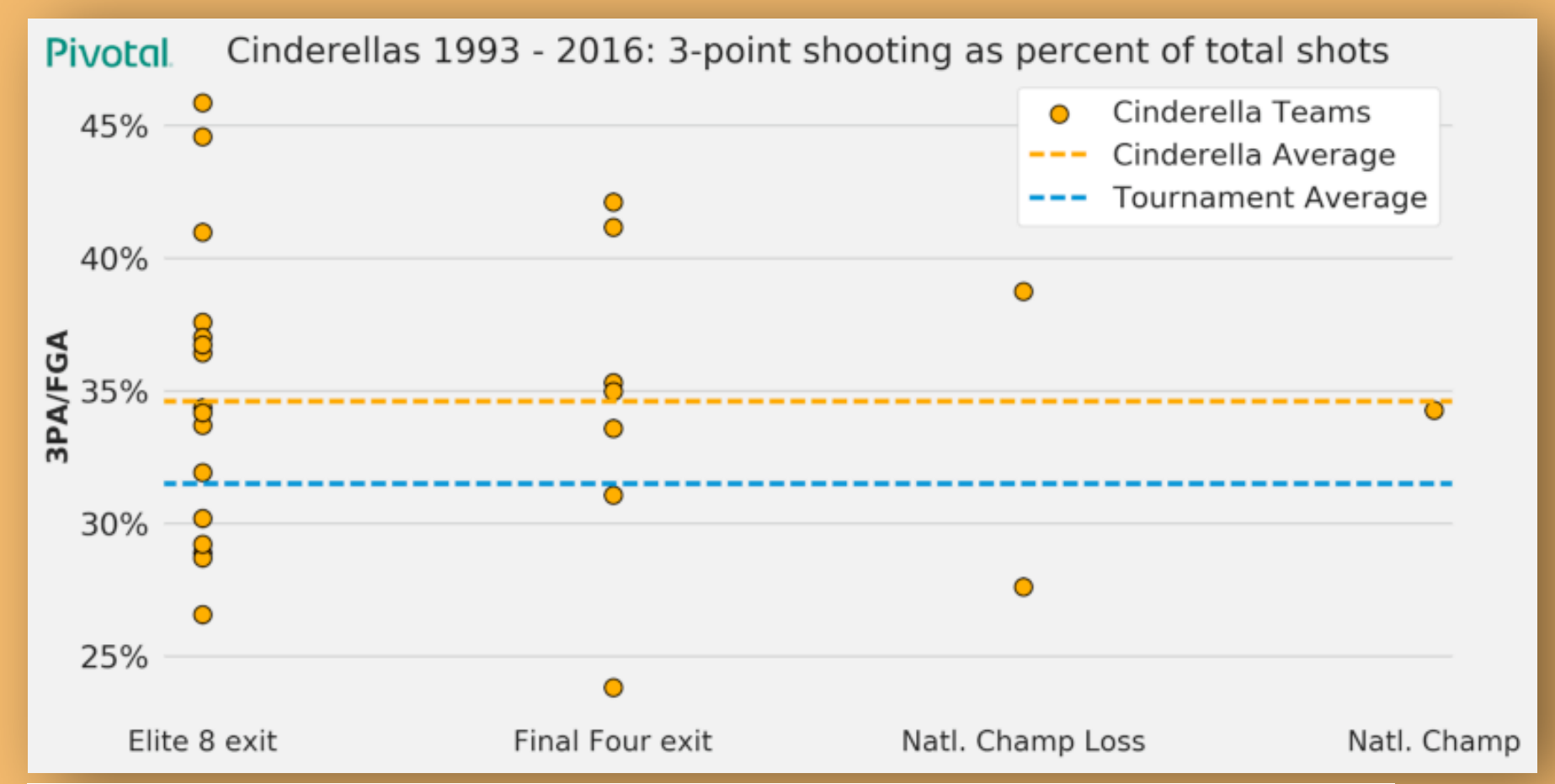
Three point attempts have been on the rise nearly every year in college basketball, with a notable exception of the 2008 season when the NCAA moved the college line out one foot to 20 feet 9 inches.

Three-point shooting: A tool for Cinderella teams in the NCAA tournament

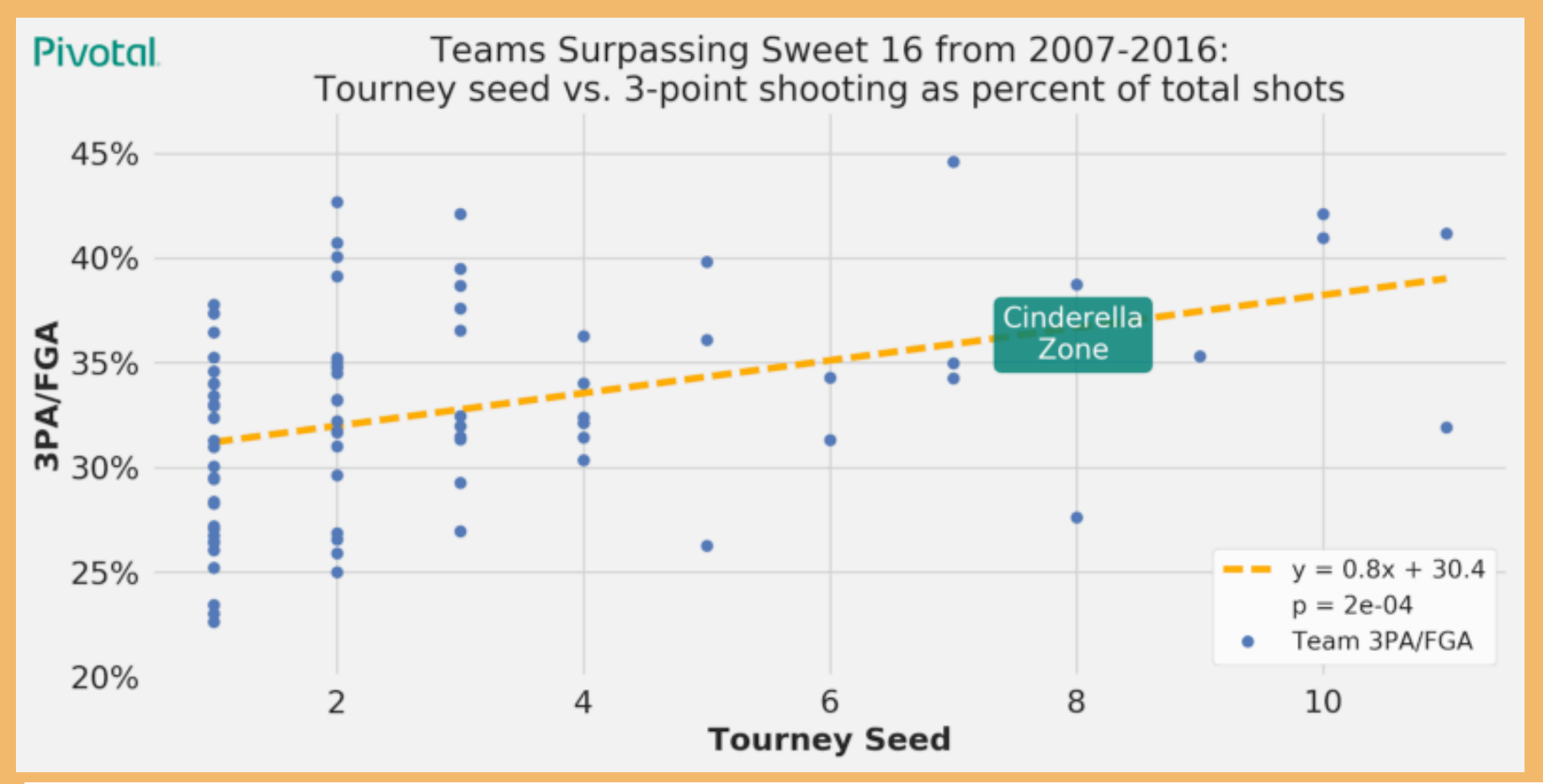
- 86 percent (166/192 teams) of teams that have made it to the Elite 8 round or better have been seeded sixth or lower, so for the sake of this analysis we've defined **teams seeded sixth or higher as "Cinderellas"**



Overview of 3-point shooting in the tournament over the last ten years along with the current AP Top 25 teams. Of the past 10 winners of the Tournament, all of them have actually attempted just under the league average per game of 3-point shots. However, collectively their 3-point accuracy (38 percent) was above average compared to the the league (36 percent).



Cinderella teams shoot 3-pointers at a higher rate compared to other tournament teams. Cinderella teams have shot an average of 35% of total shots from beyond the arc, which is 5% higher than their top seeded counterparts who shot 30% (a statistically significant result at the 1% confidence level using both a Kolmogorov Smirnov test and t-test.



For teams that have made it to the Elite 8, there is a clear correlation between seeding and 3-point shooting. In general, successful lower seeded teams rely more on the 3-point shot compared to higher seeds.

Conclusion

- Analytics have changed the landscape of basketball resulting in more 3-point attempts in both the NCAA and NBA
- Jupyter Notebook and Apache Spark provide a strong foundation for data science analysis

Additional Reading: https://builttoadapt.io/how-curry-ball-will-impact-march-madness-brackets-e2789981ba52
https://content.pivotal.io/blog/how-data-science-assists-sports