Analysis of FiveThirthyEight's Club Soccer Data

Team 3: Syed Hasan, Connor Carpenter, Orion Hunter

FiveThirtyEight (FTE) is a statistics based journal site that covers a variety of topics. As a soccer fan, I have found their Club Soccer Predictions section useful as it can shed light onto the strength of teams as well as the likelihood of results and entire seasons as a whole. They do this by assigning an *SPI* to teams. This is the central metric behind their simulations, and therefore, our analysis. In short, Soccer Power Index (SPI) is a dynamic metric created by FiveThirtyEight as a "best estimate" of a team's strength. As their website explains,

> "Given the ratings for any two teams, we can project the result of a match between them in a variety of formats — such as a league match, a home-and-away tie or a cup final — as well as simulate whole seasons to arrive at the probability each team will win the league, qualify for the Champions League or be relegated to a lower division."

The Data

A link to a github repository containing this data is published directly on their website. In terms of structure, each row of the data represents a single match between two teams and contains the following columns:

- Date - Date the match was played
- League ID - Numerical ID for league the match was played in
- League - Name of the league the match was played in
- Team 1 - Name of the home team
- Team 2 - Name of the away team
- SPI 1 - FTE's Soccer Power Index for team 1 at that time
- SPI 2 - FTE's Soccer Power Index for team 2 at that time
- Prob 1* - Estimated probability of team 1 winning
- Prob 2* - Estimated probability of team 2 winning
- Prob Tie* - Estimated probability of a draw (when applicable)
- Proj Score 1* - Projected score of team 1
- Proj Score 2* - Projected score of team 2
- Importance 1 - FTE's measure for significance of this match to team 1
- Importance 2 - FTE's measure for significance of this match to team 2
- Score 1 - Team 1's actual score
- Score 2 - Team 2's actual score
- xG 1 - Expected goals of team 1
- xG 2 - Expected goals of team 2
- NS xG 1 - Non-shot expected goals of team 1
- NS xG 2 - Non-shot expected goals of team 2
- Adj Score 1 - Team 1's adjusted score
- Adj Score 2 - Team 2's adjusted score

*Note: Columns denoted with an asterisk are measures based on calculations of the teams SPIs

Questions Raised

After looking at this data, many questions were raised.

- How do the SPI's of top leagues compare?
    - Get context as to the general strength of leagues and how they are ranked
- At what level do international competitions play at?
    - Compared to domestic leagues
- How is a single league distributed in terms of SPI?
- Can we predict a win?
    - If so, what factors are significant?

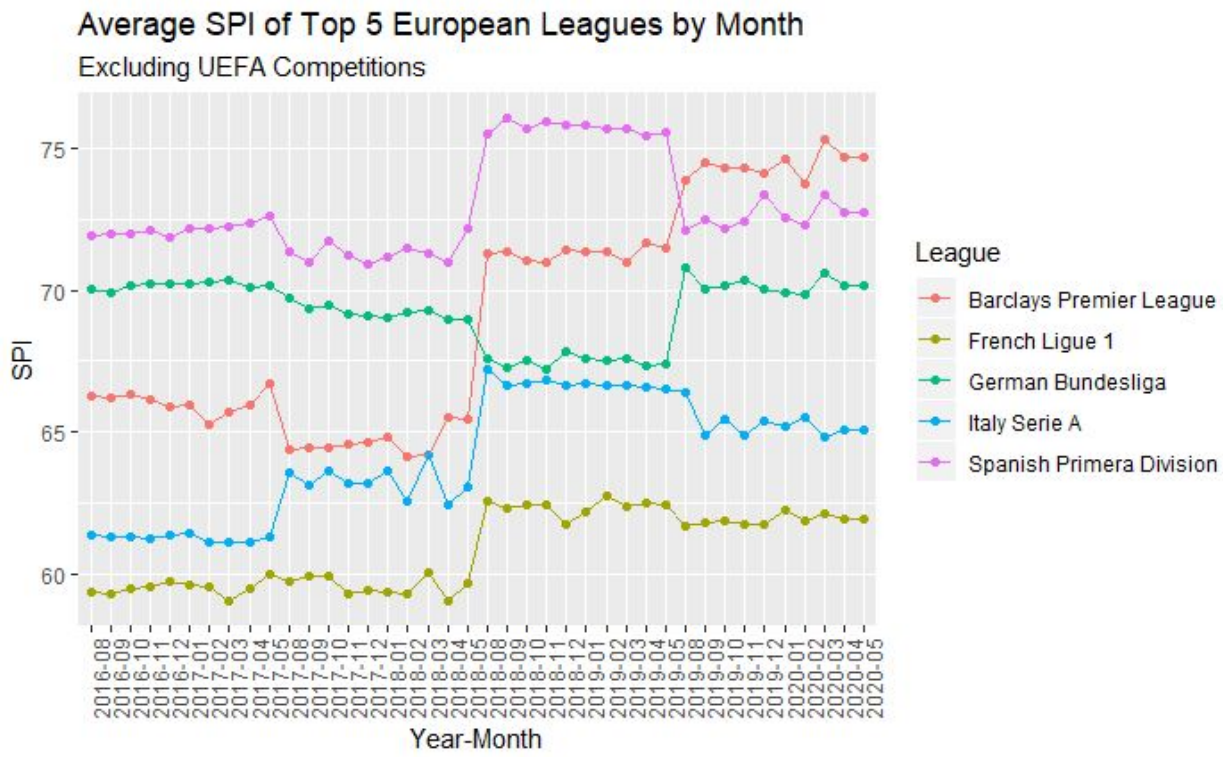
Cleaning/Transforming/Wrangling

To start cleansing our data, we will use lubridate to ensure our date column is an object of type date.

In order to answer these questions, it would be convenient to transform the data. Our goal is to make each team in each match have a row in order to more easily extract info about each team. To do this we will use cbind to attach a Match ID column so we can still track this information and then use pivot_longer to make a row for each team. In order to maintain clarity, we add a column designated whether the team is home or away, and then rename columns to suit the opponents stats. Lastly, we add a column to designate if the match was played in an international competition.

League Comparison

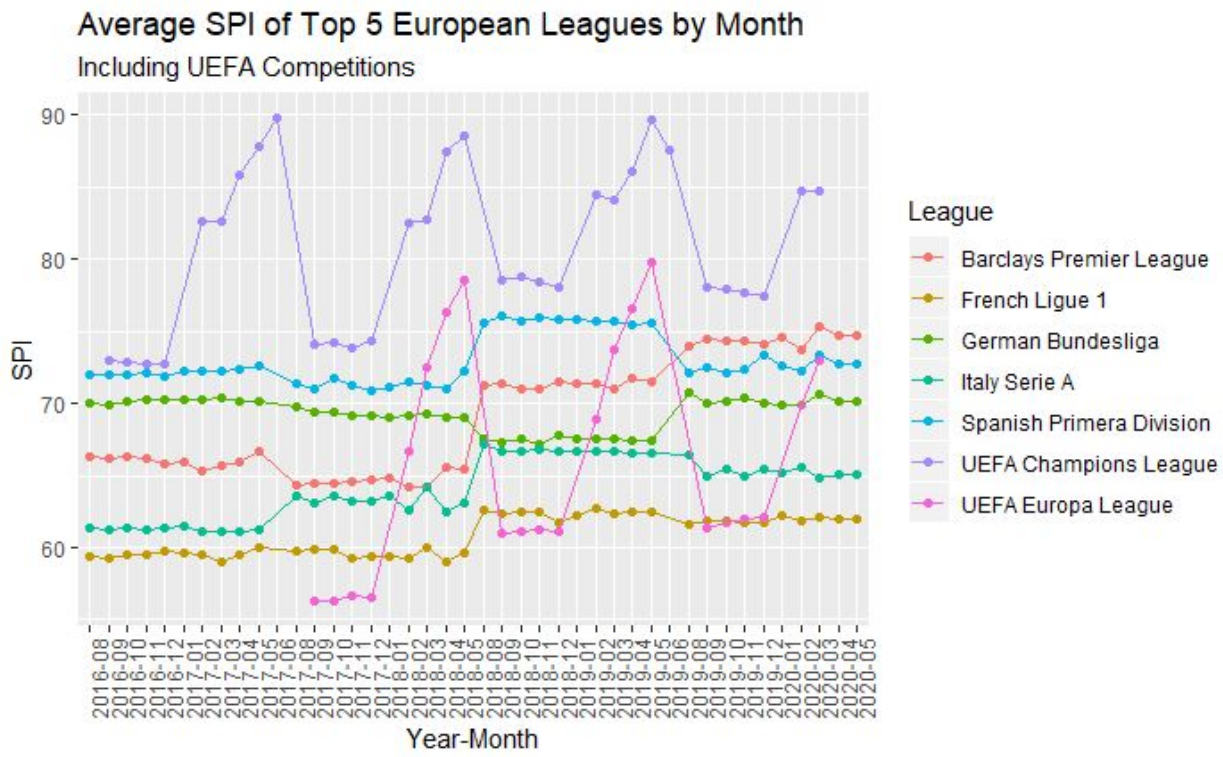
In order to resolve our initial curiosity of league strength, we must do some additional temporary wrangling. In this instance, we want to select the top 5 leagues (England, France, Germany, Italy, and Spain) and find a way to measure their SPIs over time. There are many ways to do this, but we found grouping by league and month and summarising with average SPI effective. Using ggplot with layers of geom_point and geom_line along with various formatting adjustments yields this graphic:



This gives us a good look at the average level of play in these leagues over time. Notable features in this graphic are the rapid climb of the Premier League (red line) from the 3rd overall in 2016 to the very top rated league as of the most recent ratings. Prior to the PL taking the pole position, the Spanish league had an incredibly high rated 2018-2019 season, in fact this remains the highest average SPI rating of any league in this time span. After a brief look at the SPI's in this time period, this is due to one exceptional team (Barcelona with a rating consistently above 90) along with a high congestion of teams in the 75 to 80 range.

Shortcomings of this analysis include the weakness of using mean to represent an entire league of around 20 teams. Perhaps a range could replace the points in order to give a better idea of the spread of the leagues involved. Additionally, instead of a point for each month, a better strategy could be to use the game week. This would provide for more points and a clearer view of the trajectories. Unfortunately, this was not provided in the data and could only be included with considerable effort. Lastly, since this data spans back to 2016, there are multiple seasons encapsulated in this data. It might have been helpful to somehow visualize the different seasons.

To answer our next question, we can build upon this graphic. By adding layers for the international leagues, we can compare them to the domestic leagues. This is the resulting graphic:



As these leagues are made top-division European clubs, it makes sense that they are among the upper echelon of ratings. Especially the Champions League (purple line) which is the pinnacle of european competition, repeatedly contested by only the best teams in europe. They are knockout tournaments by nature, which can be visualized by looking at the unique peaks and valleys of their lines. The peaks represent the elimination of lower strength teams and the progression of stronger teams. Then, when the new season starts the SPI returns from whence it came only for the pattern to be repeated as the season progresses.

League Analysis

In order to analyse how a league is distributed, we first had to pick one. We picked the English Premier League both because it had the highest average rating most recently, and because it is a popular league with our own personal interest.

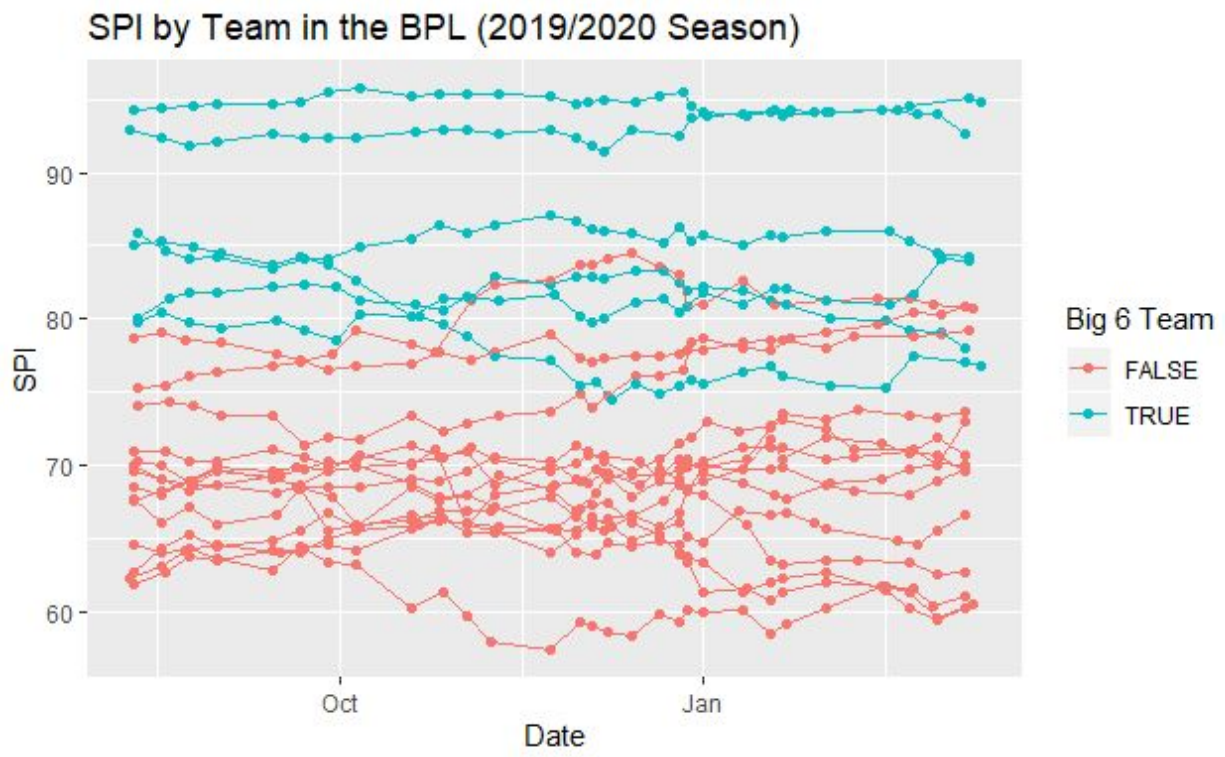
In terms of wrangling, we selected only Premier League games from the most recent season in our data. We then plotted the data the same way as before, but this time each point represented a game instead of an average over a month. This provides for a more precise graph which can be used to derive insight. Here it is:

This plot shows the SPI's for all 20 teams in the BPL. It is hard to discern much information from this as it is very crowded, however we can see that there are two dominant teams, Manchester City and Liverpool. There is a wide spread in the midfield and a very narrow spread toward the bottom, with one weak standout, Norwich City.

One drawback to this plot is its density, making it hard to draw conclusions. The solution we found for this is to break up the 20 teams into groups of interest.

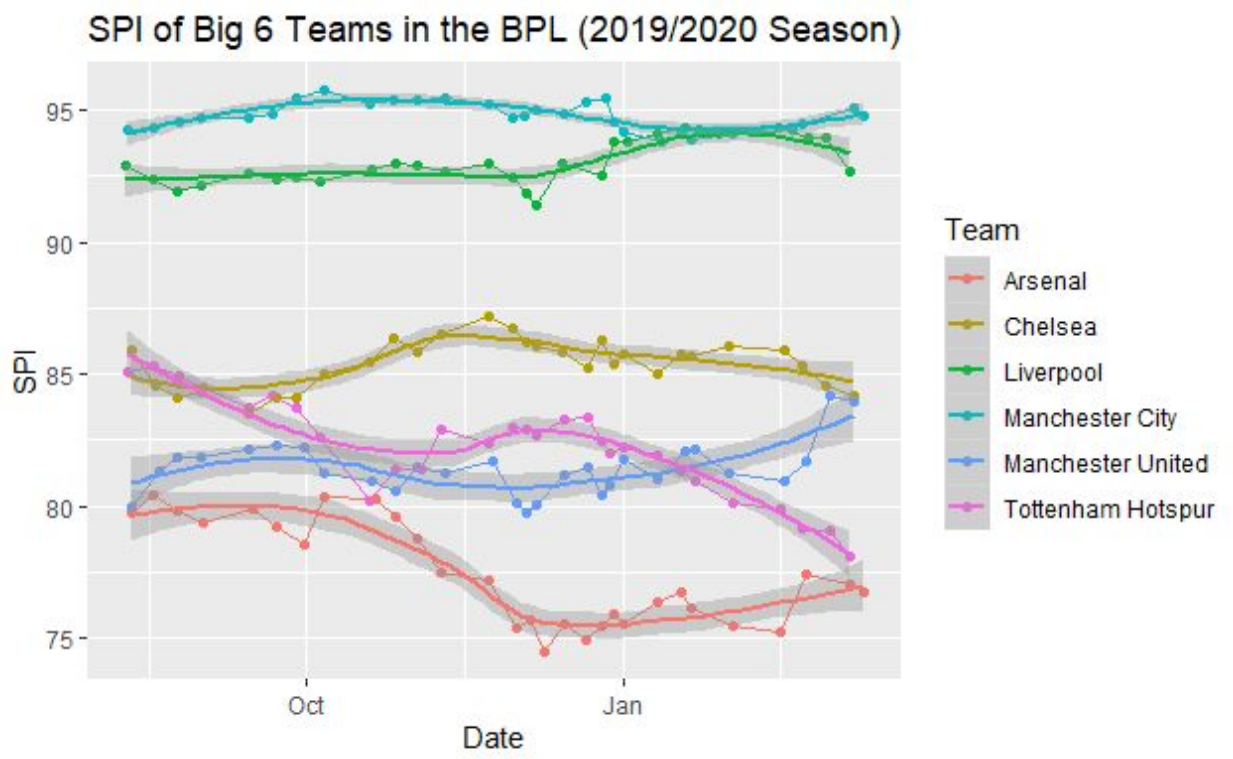
Big 6

The "Big 6" title has been given to the top six financial powerhouses in the league, usually accompanied by strong teams and good performances. These teams typically have the biggest following so it may be interesting to isolate them and look at the results. Here is a plot of these teams highlighted:



From this, we can see that while 2 of the big six teams (Manchester City and Liverpool) are dominating the field, the rest are mingling among teams outside the big six. This tells us that the big six status does not guarantee a team is among the top six teams in the league strengths. This could also be hints at a shift in the powerhouses of this league.

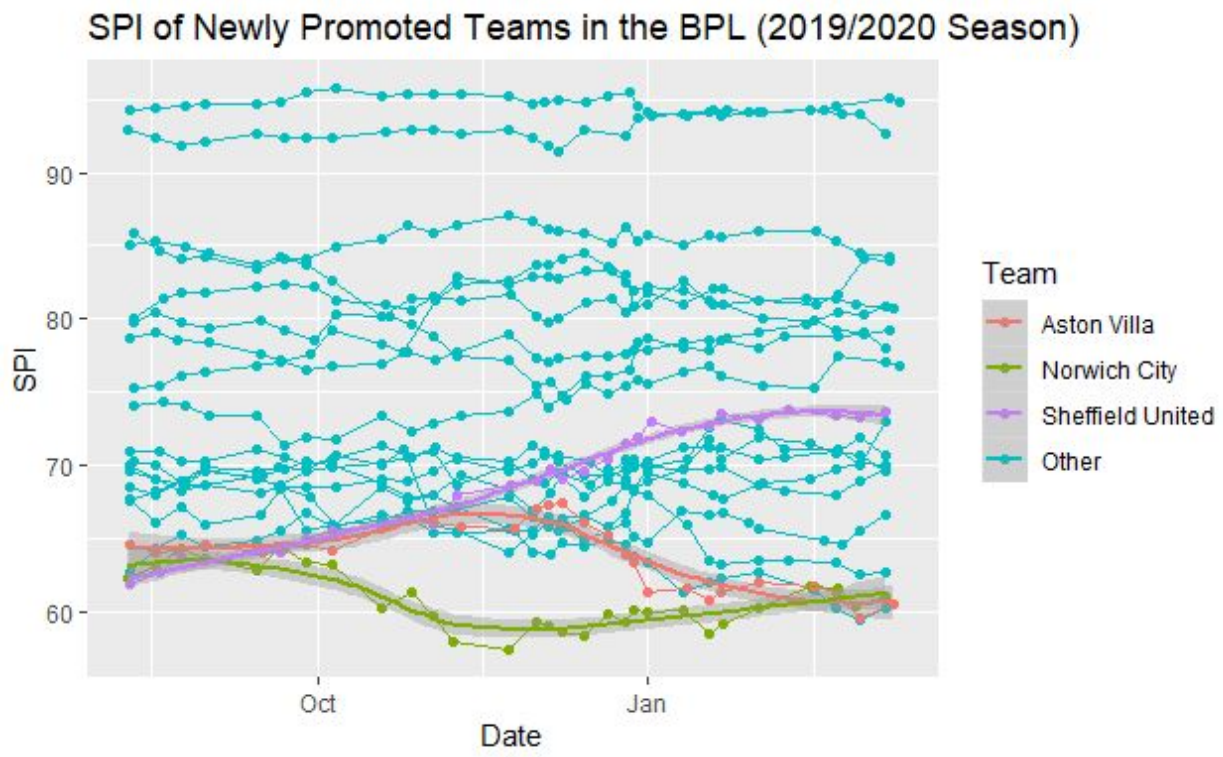
After looking at the big six compared to the league as a whole, more insights on the performances of just the big six can be had if we rid the plot of the rest of the teams. After that action is complet0, and we add a helpful geom_smooth to help visualize all six teams' trajectories, we are left with this:



This plot does very well at showing the natural ebb and flow of a team's strength. As you can see, no team stays at the same strength consistently. There are ups and downs for every single team in the plot. Notably however, is the decline of Tottenham (pink line), and the mid-season collapse or Arsenal (red line) observable from this graph. This graphic also helps visualize the gap between the top two teams and otherwise. Manchester City (aqua line) and Liverpool (green line) have maintained a gap of 5 - 7.5 SPI above other big six clubs throughout the season.

Newly Promoted Clubs

Every season, the three bottom teams are relegated to the league beneath and replaced by the top performing teams from that league. These new teams do not have high expectations, and they might make for an interesting group to look at. After combining concepts from our last few visuals, we arrive at this result:



This visual adds a geom_smooth layer exclusively over the newly promoted teams. From this we can see that Sheffield United (purple line) has had an exceptional season after being promoted from the lower league. Lastly, the other two newly promoted teams start and end near each other (both ending up declining from the start of the season), however Aston Villa (red line) had a much better season in between the start and most recent data as compared with Norwich City (green line). Despite this, other top flight teams have not proven themselves as much stronger than these teams, therefore we can not conclude who would be relegated with much confidence.

Predicting Results

In order to predict results, we must transform our data and encode factors to aid our model. To start, we must remove all null values so that our model will work. After that, we create a new column to designate a win. If a team's score is higher than the other in this match, this column will be 1, otherwise, 0. Additionally, the home and away column and the international competition are encoded similarly. After a preliminary test, it was determined that using the SPI of teams separately had little to no effect. This is likely due to the fact that what really impacts the result is the difference in SPI between teams. Based on this realization, we created two new columns; one for difference in SPI and one for difference in importance between teams. Once all of this is complete we are ready to build our model. Since we are trying to predict the result of the match, which is binary (win or tie/loss), we decided to build a logistic regression using the difference of both SPI and importance between teams, and factors of home/away and international competition as predictors. After conducting some tests against the empty/random model, we were able to conclude that this model is effective at helping predict a result. After

going through stepwise model selection, we eliminated international competition from the model because it proved insignificant in predicting results. This means that whether or not a match is played in an international competition does not impact the result. Our final model had just under 69% accuracy, which we were happy with when you consider the amount of uncertainty in sports, especially soccer which is a notoriously low scoring sport. The main flaw in the model was its sensitivity, where only around 40% of observed wins were predicted as wins.

One shortcoming to this model is the grouping of draws and losses. Perhaps a different model could help account for this but I am unaware of an alternative. Additionally, merging this data set with another to find some more variables to add to the model could prove fruitful.

I believe the significance of this model provides evidence for the accuracy of *FiveThirtyEight*'s metrics, specifically, SPI and Importance. In other words, due to the success of these metrics in being able to predict results with reasonable accuracy, these metrics must be reasonably accurate themselves, therefore we are justified in using them to compare teams.

Conclusion

Many conclusions can be had from this analysis. Here are some of the main takeaways:

- Based on the success of a logistic model to predict match results, we can conclude that FTE's SPI and Importance are accurate and important metrics
- After comparing the SPI of top domestic leagues, we can observe the rise of the Premier League
- After diving deeper into the teams of the Premier League we have two main observations:
    - Teams have a natural ebb and flow of SPI
    - On notable trends in team SPI, we can see…
        - Manchester City and Liverpool are dominating the league
        - Some Big 6 teams are not living up to expectations
        - Sheffield United's has vastly exceeded expectations after their promotion to the PL

For more information, visit the following links:
FiveThirtyEight Club Soccer Predictions
UEFA Champions League Wikipedia
Premier League Wikipedia