1. GettingAndCleaningData

March 20, 2019

```
In [1]: import time
        import requests
        from bs4 import BeautifulSoup
        import pandas as pd

        game_stats_df_list = []

        for year in range(2009,2019):

            time.sleep(.5)
            resp = requests.get("http://www.nfl.com/stats/categorystats?archive=false&conferen
                                "&role=TM&offensiveStatisticCategory=GAME_STATS&defensiveStati
                                "=null&season={0}&seasonType=REG&tabSeq=2&qualified=false&Subm

            soup = BeautifulSoup(resp.content, "html.parser")

            tables = soup.find_all("table")

            rows = []

            for team in tables[0].find_all("tr")[1:]:
                cells = team.find_all("td")

                name = cells[1].text.strip("\n")
                points_per_game = float(cells[3].text)
                total_points = int(cells[4].text.strip())
                scrimage_plays = cells[5].text.strip().replace(",", "")
                yards_per_game = float(cells[6].text.strip())
                yards_per_play = float(cells[7].text.strip())
                yards_gain_first_down = float(cells[8].text.strip())
                third_down_made = int(cells[9].text.strip())
                third_down_attempted = int(cells[10].text.strip())
                third_down_percent = int(cells[11].text.strip())
                fourth_down_made = int(cells[12].text.strip())
                fourth_down_attempted = int(cells[13].text.strip())
                fourth_down_percent = int(cells[14].text.strip())
                penalties = int(cells[15].text.strip())
                penalty_yards = cells[16].text.strip().replace(",", "")
                time_of_possession = cells[17].text.strip()
                fumbles = int(cells[18].text.strip())
                fumbles_lost = int(cells[19].text.strip())
                turnovers = int(cells[20].text.strip())
                year = year

                rows.append({
                    "Name" : name,
                    "Points per Game" : points_per_game,
                    "Total Points" : total_points,
                    "Scrimmage Plays" : scrimage_plays,
                    "Yards per Game" : yards_per_game,
                    "Yards per Play" : yards_per_play,
                    "Avg Yards Gained on 1st Down" : yards_gain_first_down,
                    "3rd Downs Made" : third_down_made,
                    "3rd Downs Attempted" : third_down_attempted,
                    "% 3rd Downs Converted" : third_down_percent,
                    "4th Downs Made" : fourth_down_made,
                    "4th Downs Attempted" : fourth_down_attempted,
                    "% 4th Downs Converted" : fourth_down_percent,
                    "Penalties" : penalties,
                    "Penalty Yards" : penalty_yards,
                    "Time of Possession" : time_of_possession,
                    "Fumbles" : fumbles,
                    "Fumbles Lost" : fumbles_lost,
                    "Turnovers" : turnovers,
                    "Year" : year
                })

            game_stats_df = pd.DataFrame(rows)
            game_stats_df = game_stats_df.set_index("Name")

            game_stats_df = game_stats_df.sort_values(by=["Year", "Total Points"])

            game_stats_df_list.append(game_stats_df)

            time.sleep(.25)

In [2]: game_stats_final_df = pd.concat(game_stats_df_list, ignore_index=False)
        game_stats_final_df;

In [3]: passing_stats_df_list = []

        for year in range(2009,2019):

            time.sleep(.5)
            resp = requests.get("http://www.nfl.com/stats/categorystats?archive=false"
                                "&conference=null&role=TM&offensiveStatisticCategory=TEAM_PASS
                                "&defensiveStatisticCategory=null&season={0}&seasonType=REG&ta
                                "=2&qualified=false&Submit=Go".format(year))

            soup = BeautifulSoup(resp.content, "html.parser")

            tables = soup.find_all("table")

            rows = []

            for team in tables[0].find_all("tr")[1:]:
                cells = team.find_all("td")

                name = cells[1].text.strip()
                pass_completions = int(cells[5].text.strip())
                pass_attempts = int(cells[6].text.strip())
                pass_completion_percentage = float(cells[7].text.strip())
                pass_attempts_per_game = float(cells[8].text.strip())
                total_pass_yards = int(cells[9].text.strip().replace(",", ""))
                passing_yards_per_game = float(cells[11].text.strip())
                passing_tds = int(cells[12].text.strip())
                interceptions = int(cells[13].text.strip())
                passes_greater_20 = int(cells[17].text.strip())
                passes_greater_40 = int(cells[18].text.strip())
                sacks = int(cells[19].text.strip())
                passer_rating = float(cells[20].text.strip())
                year = year

                rows.append({
                    "Name" : name,
                    "Pass Completions" : pass_completions,
                    "Pass Attempts" : pass_attempts,
                    "Pass Completion Percentage" : pass_completion_percentage,
                    "Pass Attempts per Game" : pass_attempts_per_game,
                    "Total Pass Yards" : total_pass_yards,
                    "Pass Yards Per Game" : passing_yards_per_game,
                    "Passing TDs" : passing_tds,
                    "Interceptions" : interceptions,
                    "Completed Passes Greater than 20 Yards" : passes_greater_20,
                    "Completed Passes Greater than 40 Yards" : passes_greater_40,
                    "Sacks" : sacks,
                    "Passer Rating" : passer_rating,
                    "Year" : year
                })

            passing_stats_df = pd.DataFrame(rows)
            passing_stats_df = passing_stats_df.set_index("Name")
```

```
            passing_stats_df_list.append(passing_stats_df)

            time.sleep(.25)

In [4]: passing_stats_final_df = pd.concat(passing_stats_df_list, ignore_index=False)
        passing_stats_final_df;

In [5]: rushing_stats_df_list = []

        for year in range(2009,2019):

            time.sleep(.5)
            resp = requests.get("http://www.nfl.com/stats/categorystats?archive=false&"
                                "conference=null&role=TM&offensiveStatisticCategory=RUSHING"
                                "&defensiveStatisticCategory=null&season={0}&seasonType=REG"
                                "&tabSeq=2&qualified=false&Submit=Go".format(year))

            soup = BeautifulSoup(resp.content, "html.parser")

            tables = soup.find_all("table")

            rows = []

            for team in tables[0].find_all("tr")[1:]:
                cells = team.find_all("td")

                name = cells[1].text.strip()
                total_rush_attempts = int(cells[5].text.strip())
                rush_attempts_per_game = float(cells[6].text.strip())
                total_rush_yards = int(cells[7].text.strip().replace(",", ""))
                rush_yards_per_carry = float(cells[8].text.strip())
                rush_yards_per_game = float(cells[9].text.strip())
                rush_TDs = int(cells[10].text.strip())
                rushes_greater_20 = int(cells[14].text.strip())
                rushes_greater_40 = int(cells[15].text.strip())
                rush_fumbles = int(cells[16].text.strip())
                year = year

                rows.append({
                    "Name" : name,
                    "Total Rush Attempts" : total_rush_attempts,
                    "Rush Attempts per Game" : rush_attempts_per_game,
                    "Total Rush Yards" : total_rush_yards,
                    "Rush Yards per Carry" : rush_yards_per_carry,
                    "Rush Yards per Game" : rush_yards_per_game,
                    "Rushing TDs" : rush_TDs,
                    "Rushes Greater than 20 Yards" : rushes_greater_20,
                    "Rushes Greater than 40 Yards" : rushes_greater_40,
                    "Rush Fumbles" : rush_fumbles,
                    "Year" : year
                })

            rushing_stats_df = pd.DataFrame(rows)
            rushing_stats_df = rushing_stats_df.set_index("Name")

            rushing_stats_df_list.append(rushing_stats_df)

            time.sleep(.25)
```

```
In [6]: rushing_stats_final_df = pd.concat(rushing_stats_df_list, ignore_index=False)
        rushing_stats_final_df;
```

```
In [7]: game_and_passing_final_df = game_stats_final_df.merge(
                        passing_stats_final_df, on=["Name", "Year"], how="inner")

        nfl = game_and_passing_final_df.merge(
                        rushing_stats_final_df, on=["Name", "Year"], how="outer")

        nfl;
```

```
In [8]: nfl.loc[0:32, "SuperBowl Winner"] = "New Orleans Saints"
        nfl.loc[32:64, "SuperBowl Winner"] = "Green Bay Packers"
        nfl.loc[64:96, "SuperBowl Winner"] = "New York Giants"
        nfl.loc[96:128, "SuperBowl Winner"] = "Baltimore Ravens"
        nfl.loc[128:160, "SuperBowl Winner"] = "Seattle Seahawks"
        nfl.loc[160:192, "SuperBowl Winner"] = "New England Patriots"
        nfl.loc[192:224, "SuperBowl Winner"] = "Denver Broncos"
        nfl.loc[224:256, "SuperBowl Winner"] = "New England Patriots"
        nfl.loc[256:288, "SuperBowl Winner"] = "Philadelphia Eagles"
        nfl.loc[288:320, "SuperBowl Winner"] = "New England Patriots"
```

```
In [9]: nfl.to_csv("nfl.csv", index=True)
```

```
In [10]: superbowl_df = nfl[((nfl.Year == 2009) & (nfl.index == "New Orleans Saints")) |
                            ((nfl.Year == 2010) & (nfl.index == "Green Bay Packers")) |
                            ((nfl.Year == 2011) & (nfl.index == "New York Giants")) |
                            ((nfl.Year == 2012) & (nfl.index == "Baltimore Ravens")) |
                            ((nfl.Year == 2013) & (nfl.index == "Seattle Seahawks")) |
                            ((nfl.Year == 2014) & (nfl.index == "New England Patriots")) |
                            ((nfl.Year == 2015) & (nfl.index == "Denver Broncos")) |
                            ((nfl.Year == 2016) & (nfl.index == "New England Patriots")) |
                            ((nfl.Year == 2017) & (nfl.index == "Philadelphia Eagles")) |
                            ((nfl.Year == 2018) & (nfl.index == "New England Patriots"))]
```

```
In [11]: superbowl_df.to_csv("superbowl_df.csv", index=True)
```

```
In [12]: worst_team_df = nfl[((nfl.Year == 2009) & (nfl.index == "St. Louis Rams")) |
                         ((nfl.Year == 2010) & (nfl.index == "Carolina Panthers")) |
                         ((nfl.Year == 2011) & (nfl.index == "Indianapolis Colts")) |
                         ((nfl.Year == 2012) & (nfl.index == "Jacksonville Jaguars")) |
                         ((nfl.Year == 2013) & (nfl.index == "Houston Texans")) |
                         ((nfl.Year == 2014) & (nfl.index == "Tampa Bay Buccaneers")) |
                         ((nfl.Year == 2015) & (nfl.index == "Tennessee Titans")) |
                         ((nfl.Year == 2016) & (nfl.index == "Cleveland Browns")) |
                         ((nfl.Year == 2017) & (nfl.index == "Cleveland Browns")) |
                         ((nfl.Year == 2018) & (nfl.index == "Arizona Cardinals"))]
```

```
In [13]: worst_team_df.to_csv("worst_team_df.csv", index=True)
```