College Football Data Collection and Preprocessing

Group Members

Caleb Miller: cm3962@drexel.edu

Muhammad Hazrat: mh3852@drexel.edu

Hashim Afzal: ha695@drexel.edu

DSCI 511 – 003

Dr. Shadi Rezapour

Introduction

- This dataset offers a comprehensive view of college football games and teams
- This includes scoring details, team information, excitement index, and attendance figures
- The data collected is from the 2022 college football season

Purpose of the Dataset

- The purpose of this dataset is to analyze factors influencing attendance and excitement in college football games
  - High scoring vs low scoring
  - Close game vs blowout
  - Level of competition (D1, D2, etc.)
- The dataset allows the extraction of valuable insights for sports administrators
- This can be helpful for strategic planning and enhancing fan experiences

Potential Users and Applications

- Sports Analytics: The dataset can be utilized for in-depth analysis of game trends and team performance
- Sports Administration: Insights from data can be leveraged to make strategic decisions such as:
  - Game scheduling
  - Fan engagement
  - Marketing strategies
- Academic Research: Helps understand the impact of college sports on student life and local economies (more data collection needed)

Data Source and Acquisition

- The data was acquired from collegefootballdata.com
  - A comprehensive source for college football statistics and information
- Obtained using an API Key
  - Allows access to data covering different aspects of games such as:
    - Team Statistics
    - Player Statistics
    - Game Statistics
- The API allows users to extract large volumes of data efficiently

Data Preprocessing

- Formatting:
  - The timestamp had to be reformatted into a usable form (date and time as separate variables)
- Further extraction:
  - Box scores for home and away teams had to be separated into Q1, Q2, Q3, and Q4
  - Team records for total, conference, home, and away games extracted to their own column

date	start_time
8/27/22	00:00.0
8/27/22	00:00.0
8/27/22	00:00.0
8/27/22	00:00.0
8/27/22	00:00.0
8/28/22	15:00.0
8/28/22	00:00.0
8/28/22	30:00.0
9/1/22	00:00.0

```
pattern = r"(\d{4}-\d{2}-\d{2})T(\d{2}:\d{2}:\d{2}.\d{3})Z"

# Use re.search to find the pattern in the text

df[['date', 'start_time']] = df['start_date'].str.extract(pattern)
```

Data Preprocessing

- Handling missing data:
  - ‘Pre/Postgame elo’ and ‘Post Win Probability’ for home and away teams were removed due to missing entries
  - The dataset was refined to only include rows that are not missing ‘Attendance’ and ‘Excitement Index’
    - These variables are necessary

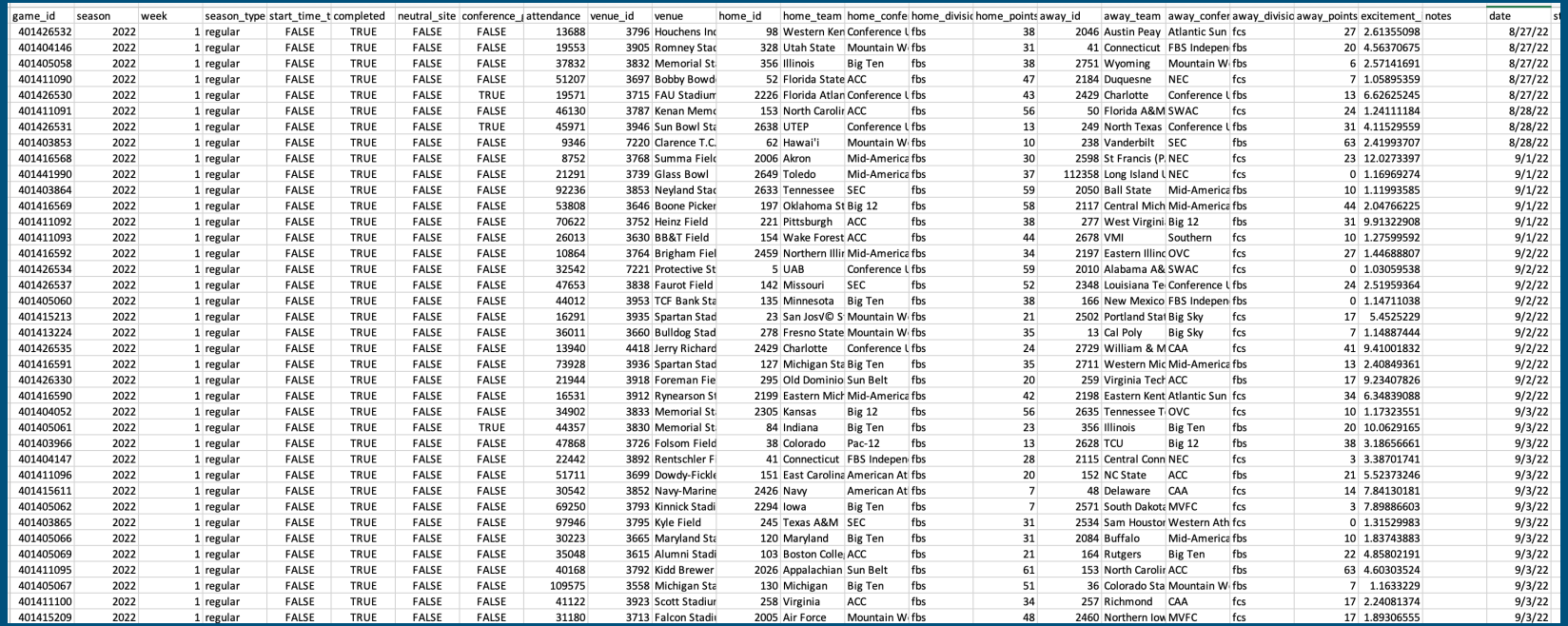
```
df.drop('away_postgame_elo', axis=1, inplace=True)
df.drop('home_postgame_elo', axis=1, inplace=True)
df.drop('away_pregame_elo', axis=1, inplace=True)
df.drop('home_pregame_elo', axis=1, inplace=True)
df.drop('away_post_win_prob', axis=1, inplace=True)
df.drop('home_post_win_prob', axis=1, inplace=True)

df = df.dropna(subset=['attendance'])

df = df.dropna(subset=['excitement_index'])
```

year	team_id	team	conference	division	expected_wi	total_games	total_wins	total_losses	total_ties	total_confer	conference_v	conference_l	conference_t	total_home_	home_wins	home_losses	home_ties	total_away_	away_wins	away_losses	away_ties
2022	2	Auburn	SEC	West	6.7	12	5	7	0	8	2	6	0	8	5	3	0	4	0	4	0
2022	5	UAB	Conference USA		8.6	13	7	6	0	8	4	4	0	6	5	1	0	6	1	5	0
2022	6	South Alabar	Sun Belt	West	10.3	13	10	3	0	8	7	1	0	6	5	1	0	6	5	1	0
2022	8	Arkansas	SEC	West	8.5	13	7	6	0	8	3	5	0	7	4	3	0	4	2	2	0
2022	9	Arizona Stat	Pac-12		4.3	12	3	9	0	9	2	7	0	6	2	4	0	6	1	5	0
2022	12	Arizona	Pac-12		4	12	5	7	0	9	3	6	0	7	3	4	0	5	2	3	0
2022	21	San Diego St	Mountain W	West	7.6	13	7	6	0	8	5	3	0	7	5	2	0	5	2	3	0
2022	23	San JosV© S	Mountain W	West	5.5	12	7	5	0	8	5	3	0	6	6	0	0	5	1	4	0
2022	24	Stanford	Pac-12		2.9	12	3	9	0	9	1	8	0	6	2	4	0	6	1	5	0
2022	25	California	Pac-12		3.7	12	4	8	0	9	2	7	0	7	4	3	0	5	0	5	0
2022	26	UCLA	Pac-12		9.8	13	9	4	0	9	6	3	0	8	6	2	0	4	3	1	0
2022	30	USC	Pac-12		9.8	14	11	3	0	10	8	2	0	7	7	0	0	5	4	1	0
2022	36	Colorado Sta	Mountain W	Mountain	4	12	3	9	0	8	3	5	0	6	2	4	0	6	1	5	0
2022	38	Colorado	Pac-12		1	12	1	11	0	9	1	8	0	6	1	5	0	6	0	6	0
2022	41	Connecticut	FBS Independents		6.3	13	6	7	0	0	0	0	0	6	5	1	0	6	1	5	0
2022	52	Florida State	ACC	Atlantic	9.5	13	10	3	0	8	5	3	0	7	5	2	0	4	3	1	0
2022	57	Florida	SEC	East	6.3	13	6	7	0	8	3	5	0	7	5	2	0	4	1	3	0
2022	58	South Florid	American Athletic		3.2	12	1	11	0	8	0	8	0	5	1	4	0	6	0	6	0
2022	59	Georgia Tech	ACC	Coastal	3.7	12	5	7	0	8	4	4	0	5	2	3	0	6	3	3	0
2022	61	Georgia	SEC	East	14.4	15	15	0	0	9	9	0	0	6	6	0	0	4	4	0	0
2022	62	Hawai'i	Mountain W	West	2.9	13	3	10	0	8	2	6	0	7	3	4	0	6	0	6	0
2022	66	Iowa State	Big 12		4.7	12	4	8	0	9	1	8	0	7	3	4	0	5	1	4	0
2022	68	Boise State	Mountain W	Mountain	8.6	14	10	4	0	9	8	1	0	7	5	2	0	6	4	2	0
2022	77	Northwester	Big Ten	West	3.5	12	1	11	0	9	1	8	0	6	0	6	0	5	0	5	0
2022	84	Indiana	Big Ten	East	1.7	12	4	8	0	9	2	7	0	7	3	4	0	5	1	4	0
2022	87	Notre Dame	FBS Independents		8.7	13	9	4	0	1	1	0	0	6	4	2	0	4	2	2	0
2022	96	Kentucky	SEC	East	7.2	13	7	6	0	8	3	5	0	8	5	3	0	4	2	2	0
2022	97	Louisville	ACC	Atlantic	8.3	13	8	5	0	8	4	4	0	6	5	1	0	6	2	4	0
2022	98	Western Ken	Conference USA		8.6	14	9	5	0	8	6	2	0	6	4	2	0	7	4	3	0
2022	99	LSU	SEC	West	8.9	14	10	4	0	9	6	3	0	7	6	1	0	4	3	1	0
2022	103	Boston Colle	ACC	Atlantic	3.1	12	3	9	0	8	2	6	0	6	2	4	0	6	1	5	0
2022	113	UMass	FBS Independents		1.7	12	1	11	0	3	0	3	0	5	1	4	0	7	0	7	0
2022	120	Maryland	Big Ten	East	6.8	13	8	5	0	9	4	5	0	7	5	2	0	5	2	3	0
2022	127	Michigan Sta	Big Ten	East	3.9	12	5	7	0	9	3	6	0	7	4	3	0	5	1	4	0
2022	130	Michigan	Big Ten	East	12.6	14	13	1	0	10	10	0	0	8	8	0	0	4	4	0	0
2022	135	Minnesota	Big Ten	West	9.7	13	9	4	0	9	5	4	0	7	5	2	0	5	3	2	0

TeamsData.csv



game_id	season	week	season_type	start_time_t	completed	neutral_site	conference_	attendance	venue_id	venue	home_id	home_team	home_confe	home_divisi	home_points	away_id	away_team	away_confer	away_divisi	away_points	excitement_	notes	date	st
401426532	2022	1	regular	FALSE	TRUE	FALSE	FALSE	13688	3796	Houchens Ind	98	Western Ken	Conference U	fbs	38	2046	Austin Peay	Atlantic Sun	fcs	27	2.61355098		8/27/22	
401404146	2022	1	regular	FALSE	TRUE	FALSE	FALSE	19553	3905	Romney Stac	328	Utah State	Mountain W	fbs	31	41	Connecticut	FBS Indepen	fbs	20	4.56370675		8/27/22	
401405058	2022	1	regular	FALSE	TRUE	FALSE	FALSE	37832	3832	Memorial St	356	Illinois	Big Ten	fbs	38	2751	Wyoming	Mountain W	fbs	6	2.57141691		8/27/22	
401411090	2022	1	regular	FALSE	TRUE	FALSE	FALSE	51207	3697	Bobby Bowd	52	Florida State	ACC	fbs	47	2184	Duquesne	NEC	fcs	7	1.05895359		8/27/22	
401426530	2022	1	regular	FALSE	TRUE	FALSE	TRUE	19571	3715	FAU Stadium	2226	Florida Atlan	Conference U	fbs	43	2429	Charlotte	Conference U	fbs	13	6.62625245		8/27/22	
401411091	2022	1	regular	FALSE	TRUE	FALSE	FALSE	46130	3787	Kenan Memo	153	North Carolir	ACC	fbs	56	50	Florida A&M	SWAC	fcs	24	1.24111184		8/28/22	
401426531	2022	1	regular	FALSE	TRUE	FALSE	TRUE	45971	3946	Sun Bowl Sta	2638	UTEP	Conference U	fbs	13	249	North Texas	Conference U	fbs	31	4.11529559		8/28/22	
401403853	2022	1	regular	FALSE	TRUE	FALSE	FALSE	9346	7220	Clarence T.C.	62	Hawai'i	Mountain W	fbs	10	238	Vanderbilt	SEC	fbs	63	2.41993707		8/28/22	
401416568	2022	1	regular	FALSE	TRUE	FALSE	FALSE	8752	3768	Summa Field	2006	Akron	Mid-America	fbs	30	2598	St Francis (P	NEC	fcs	23	12.0273397		9/1/22	
401441990	2022	1	regular	FALSE	TRUE	FALSE	FALSE	21291	3739	Glass Bowl	2649	Toledo	Mid-America	fbs	37	112358	Long Island U	NEC	fcs	0	1.16969274		9/1/22	
401403864	2022	1	regular	FALSE	TRUE	FALSE	FALSE	92236	3853	Neyland Stac	2633	Tennessee	SEC	fbs	59	2050	Ball State	Mid-America	fbs	10	1.11993585		9/1/22	
401416569	2022	1	regular	FALSE	TRUE	FALSE	FALSE	53808	3646	Boone Picker	197	Oklahoma St	Big 12	fbs	58	2117	Central Mich	Mid-America	fbs	44	2.04766225		9/1/22	
401411092	2022	1	regular	FALSE	TRUE	FALSE	FALSE	70622	3752	Heinz Field	221	Pittsburgh	ACC	fbs	38	277	West Virgini	Big 12	fbs	31	9.91322908		9/1/22	
401411093	2022	1	regular	FALSE	TRUE	FALSE	FALSE	26013	3630	BB&T Field	154	Wake Forest	ACC	fbs	44	2678	VMI	Southern	fcs	10	1.27599592		9/1/22	
401416592	2022	1	regular	FALSE	TRUE	FALSE	FALSE	10864	3764	Brigham Fiel	2459	Northern Illir	Mid-America	fbs	34	2197	Eastern Illinc	OVC	fcs	27	1.44688807		9/2/22	
401426534	2022	1	regular	FALSE	TRUE	FALSE	FALSE	32542	7221	Protective St	5	UAB	Conference U	fbs	59	2010	Alabama A&	SWAC	fcs	0	1.03059538		9/2/22	
401426537	2022	1	regular	FALSE	TRUE	FALSE	FALSE	47653	3838	Faurot Field	142	Missouri	SEC	fbs	52	2348	Louisiana Te	Conference U	fbs	24	2.51959364		9/2/22	
401405060	2022	1	regular	FALSE	TRUE	FALSE	FALSE	44012	3953	TCF Bank Sta	135	Minnesota	Big Ten	fbs	38	166	New Mexico	FBS Indepen	fbs	0	1.14711038		9/2/22	
401415213	2022	1	regular	FALSE	TRUE	FALSE	FALSE	16291	3935	Spartan Stad	23	San Josv© S	Mountain W	fbs	21	2502	Portland Stat	Big Sky	fcs	17	5.4525229		9/2/22	
401413224	2022	1	regular	FALSE	TRUE	FALSE	FALSE	36011	3660	Bulldog Stad	278	Fresno State	Mountain W	fbs	35	13	Cal Poly	Big Sky	fcs	7	1.14887444		9/2/22	
401426535	2022	1	regular	FALSE	TRUE	FALSE	FALSE	13940	4418	Jerry Richard	2429	Charlotte	Conference U	fbs	24	2729	William & M	CAA	fcs	41	9.41001832		9/2/22	
401416591	2022	1	regular	FALSE	TRUE	FALSE	FALSE	73928	3936	Spartan Stad	127	Michigan Sta	Big Ten	fbs	35	2711	Western Mic	Mid-America	fbs	13	2.40849361		9/2/22	
401426330	2022	1	regular	FALSE	TRUE	FALSE	FALSE	21944	3918	Foreman Fie	295	Old Dominio	Sun Belt	fbs	20	259	Virginia Tech	ACC	fbs	17	9.23407826		9/2/22	
401416590	2022	1	regular	FALSE	TRUE	FALSE	FALSE	16531	3912	Rynearson St	2199	Eastern Mich	Mid-America	fbs	42	2198	Eastern Kent	Atlantic Sun	fcs	34	6.34839088		9/2/22	
401404052	2022	1	regular	FALSE	TRUE	FALSE	FALSE	34902	3833	Memorial St	2305	Kansas	Big 12	fbs	56	2635	Tennessee T	OVC	fcs	10	1.17323551		9/3/22	
401405061	2022	1	regular	FALSE	TRUE	FALSE	TRUE	44357	3830	Memorial St	84	Indiana	Big Ten	fbs	23	356	Illinois	Big Ten	fbs	20	10.0629165		9/3/22	
401403966	2022	1	regular	FALSE	TRUE	FALSE	FALSE	47868	3726	Folsom Field	38	Colorado	Pac-12	fbs	13	2628	TCU	Big 12	fbs	38	3.18656661		9/3/22	
401404147	2022	1	regular	FALSE	TRUE	FALSE	FALSE	22442	3892	Rentschler F	41	Connecticut	FBS Indepen	fbs	28	2115	Central Conn	NEC	fcs	3	3.38701741		9/3/22	
401411096	2022	1	regular	FALSE	TRUE	FALSE	FALSE	51711	3699	Dowdy-Ficke	151	East Carolina	American At	fbs	20	152	NC State	ACC	fbs	21	5.52373246		9/3/22	
401415611	2022	1	regular	FALSE	TRUE	FALSE	FALSE	30542	3852	Navy-Marine	2426	Navy	American At	fbs	7	48	Delaware	CAA	fcs	14	7.84130181		9/3/22	
401405062	2022	1	regular	FALSE	TRUE	FALSE	FALSE	69250	3793	Kinnick Stadi	2294	Iowa	Big Ten	fbs	7	2571	South Dakot	MVFC	fcs	3	7.89886603		9/3/22	
401403865	2022	1	regular	FALSE	TRUE	FALSE	FALSE	97946	3795	Kyle Field	245	Texas A&M	SEC	fbs	31	2534	Sam Houstor	Western Ath	fcs	0	1.31529983		9/3/22	
401405066	2022	1	regular	FALSE	TRUE	FALSE	FALSE	30223	3665	Maryland Sta	120	Maryland	Big Ten	fbs	31	2084	Buffalo	Mid-America	fbs	10	1.83743883		9/3/22	
401405069	2022	1	regular	FALSE	TRUE	FALSE	FALSE	35048	3615	Alumni Stadi	103	Boston Colle	ACC	fbs	21	164	Rutgers	Big Ten	fbs	22	4.85802191		9/3/22	
401411095	2022	1	regular	FALSE	TRUE	FALSE	FALSE	40168	3792	Kidd Brewer	2026	Appalachian	Sun Belt	fbs	61	153	North Carolir	ACC	fbs	63	4.60303524		9/3/22	
401405067	2022	1	regular	FALSE	TRUE	FALSE	FALSE	109575	3558	Michigan Sta	130	Michigan	Big Ten	fbs	51	36	Colorado Sta	Mountain W	fbs	7	1.1633229		9/3/22	
401411100	2022	1	regular	FALSE	TRUE	FALSE	FALSE	41122	3923	Scott Stadiur	258	Virginia	ACC	fbs	34	257	Richmond	CAA	fcs	17	2.24081374		9/3/22	
401415209	2022	1	regular	FALSE	TRUE	FALSE	FALSE	31180	3713	Falcon Stadi	2005	Air Force	Mountain W	fbs	48	2460	Northern Iow	MVFC	fcs	17	1.89306555		9/3/22	

GamesData.csv

Distribution Approach

- We aim to maximize accessibility and usability of the dataset
- The data is provided in two CSV files:
  - ‘GamesData.csv’ contains each individual game as a data row
  - ‘TeamsData.csv’ contains each team as a data row
  - These files are meant to be used in conjunction to derive deeper insights
- Accessibility is enhanced by providing detailed documentation (README file) that helps users navigate and use the data effectively

Access Rights

- The dataset is available for educational, research, and analytical purposes
  - This promotes widespread use in these communities
- Users are required to agree to the terms and conditions set by collegefootballdata.com
  - This includes restrictions on commercial use and redistribution
    - Commercial entities may require special permissions/licenses for uses involving monetization

Issues and Limitations

- Absence of key variables: the lack of key variables impact the extent of available insights
  - Team rankings
  - Revenue data
- Incomplete records: missing data limits the accuracy of any analysis
  - Specifically, dealing with missing ‘Attendance’ and ‘Excitement Index’ entries caused the dataset to be much smaller
  - The ‘Notes’ variable has missing entries, but this is acceptable
    - Any entries are special occasions

Team and Contributions

- Team members:
  - Caleb Miller
  - Hashim Afzal
  - Muhammad Hazrat
- Our group worked collectively on all stages of the project
- Data collection and preprocessing was lead by Caleb and Hashim
- The README documentation was lead by Muhammad
- The presentation was completed as a whole

Thank you!

- Questions?
- Feel free to email us:
  - Caleb Miller: cm3962@drexel.edu
  - Muhammad Hazrat: mh3852@drexel.edu
  - Hashim Afzal: ha695@drexel.edu