BUDT 703 DBMS Readme document

1. Data Source

- The below link gives the information about the opponent, date, time, tournament, venue, games scores and game result. This data is available for every year starting 1999. And we use the text only option to generate the data in an extractable format.
  https://umterps.com/sports/baseball/schedule

Date	Time	At	Opponent	Location	Tournament	Result
Feb 18 (Fri)	7:30	Away	Baylor	Waco, TX		W 4-0
Feb 19 (Sat)	4:00	Away	Baylor	Waco, TX		W 9-5
Feb 20 (Sun)	2:00	Away	Baylor	Waco, TX		W 8-4
Feb 23 (Wed)	4:00	Home	UMBC	College Park, MD		W 3-2

- Along with this we used google to get the venue name. For example, if Terps play UMBC at an away game in UMBC then the venue name is Baseball Factory.

- Similarly, we added the university name for the respective opponent team.

2. Data cleaning and pre-processing

- We paste the data into excel from the page shown above.

	A	B	C	D	E	F	G	H
1	Date	Time	At	Opponent	Location	Tournament Resu	Year	Tournament
2	Jan 29 (Fri)	3:00 PM	Home	Oklahoma	ACC-Disney Blast Orlando (vs. Oklahoma)	L 9-10	1999	Club Friendlies
3	Jan 30 (Sat)	7:00 PM	Home	Jacksonville	ACC-Disney Blast Orlando (vs. Jacksonville)	L 2-12	1999	Club Friendlies
4	Jan 31 (Sun)	3:00 PM	Home	Auburn	ACC-Disney Blast Orlando (vs. Auburn)		1999	Club Friendlies
5	Feb 12 (Fri)	3:00 PM	Away	No Carolina A&T	North Carolina AT (Rained out)		1999	Club Friendlies
6	Feb 13 (Sat)	12:00 PM	Away	Elon	Elon College	W 14-3	1999	Club Friendlies
7	Feb 13 (Sat)	3:00 PM	Away	Elon	Elon College	W 8-3	1999	Club Friendlies
8	Feb 14 (Sun)	1:00 PM	Away	Elon	Elon College	W 3-2	1999	Club Friendlies
9	Feb 20 (Sat)	1:00 PM	Away	Nc Greensboro	UNC-Greensboro		1999	Club Friendlies
10	Feb 20 (Sat)	4:00 PM	Away	Nc Greensboro	UNC-Greensboro		1999	Club Friendlies
11	Feb 21 (Sun)	1:00 PM	Away	Nc Greensboro	UNC-Greensboro	L 4-7	1999	Club Friendlies

- Opponent Data
  We assign each unique opponent a unique ID, and then we have the opponent Name and Opponent university. We then use the adjacent cell to concatenate the values as per the format required in the INSERT SQL query as highlighted in the top. While doing this we ensured different variations of the same opponent name is normalized to one name and the ID is allocated only once.

	A	B	C	D
1	INSERT INTO [DataBase.Opponent](opponentId, opponentName, opponentUniversity) VALUES			
2	opponentID	opponentName	opponentUniversity	('opponentId','opponentName','opponentUniversity')
3	O001	Oklahoma	University of Oklahoma	('O001','Oklahoma','University of Oklahoma'),
4	O002	Jacksonville	Jacksonville University	('O002','Jacksonville','Jacksonville University'),
5	O003	Auburn	Auburn University	('O003','Auburn','Auburn University'),
6	O004	North Carolina A&T	North Carolina A&T University	('O004','North Carolina A&T','North Carolina A&T University'),
7	O005	Elon	Elon University	('O005','Elon','Elon University'),
8	O006	Nc Greensboro	University of Nc Greensboro	('O006','Nc Greensboro','University of Nc Greensboro'),

- Tournament Data
  We assign each unique Tournament a unique ID, and then we have the tournament Name. We then use the adjacent cell to concatenate the values as per the format required in the INSERT SQL query as highlighted in the top.
  Since Tournament name couldn't be extracted from the text page of the UMD baseball schedule website, we manually added the data from the main schedule page for every year.

A	B	C
INSERT INTO [DataBase.Tournament] (tournamentId, tournamentName) VALUES		
tournamentId	tournamentName	('tournamentId','tournamentName'),
T01	Club Friendlies	('T01','Club Friendlies'),
T02	Big Ten	('T02','Big Ten'),
T03	ACC Championship	('T03','ACC Championship'),
T04	NCAA Tournament	('T04','NCAA Tournament')

- Calendar Data
  Here we have the list of months and year We then use the adjacent cell to concatenate the values as per the format required in the INSERT SQL query as highlighted in the top.

A	B	C	D	E
INSERT INTO [DataBase.Calendar] (calendarMonth, calendarYear) VALUES				
calendarMonth	calendarYear	('calendarMonth','calendarYear'),		
Jan	2023	('Jan','2023'),		
Feb	2023	('Feb','2023'),		
Mar	2023	('Mar','2023'),		
Apr	2023	('Apr','2023'),		
May	2023	('May','2023'),		

- Venue Data
  We assign each unique Venue a unique ID, and then we have the venue Name, venue city and venue state. We then use the adjacent cell to concatenate the values as per the format required in the INSERT SQL query as highlighted in the top. While doing this we ensured different variations of the same venue is normalized to one name and the ID is allocated only once.

A	B	C	D	E
INSERT INTO [DataBase.Venue](venueId, venueName,venueLocation, -venueCity, -venueState) VALUES				
venueId	**venueName**	**venueCity**	**venueSt**	**('venueId','venueName','venueCity','venueState'),**
V001	USF Baseball Stadium	Tampa	FL	('V001','USF Baseball Stadium','Tampa','FL'),
V002	Bob "Turtle" Smith Stadium	College Park	MD	('V002','Bob "Turtle" Smith Stadium','College Park','MD'),
V003	Swayze Field	Oxford	MS	('V003','Swayze Field','Oxford','MS'),
V004	U.S. Bank Stadium	Minneapolis	MN	('V004','U.S. Bank Stadium','Minneapolis','MN'),
V005	Bob Hanna Stadium	Newark	DE	('V005','Bob Hanna Stadium','Newark','DE'),
V006	John Euliano Park	Orlando	FL	('V006','John Euliano Park','Orlando','FL'),
V007	Duane Banks Field	Iowa	IA	('V007','Duane Banks Field','Iowa','IA'),
V008	O`Brate Stadium	Columbus	OH	('V008','O`Brate Stadium','Columbus','OH'),

- Play Data

The data being reported in the table are as follows, date, time, **month, year, opponent ID, tournament ID, Venue ID**, Game type (Home or Away), Terps' score and opponent score. The items bolded are foreign keys and are filled in Excel using VLOOKUP. We then use the adjacent cell to concatenate the values as per the format required in the INSERT SQL query as highlighted in the top.

A	B	C	D	E	F	G	H	I	J	K
INSERT INTO [DataBase.Play]Play (playTime,playDate,calendarMonth,calendarYear,opponentId, tournamentId, venueId,playType, playTerpScore, playOpponentSc										
Play (playTime,playDate,calendarMonth,calendarYear,opponentId, tournamentId, venueId,playType, playTerpScore, playOpponentScore)										
playTime	playDate	caler	calendarY	opponentId	tourna	venueId	playType	playTe	playOppo	
3:00 PM	29	Jan	1999	O001	T01	V016	Home	9	10	('03:00 PM','28','Jan','1999','O001','T01','V016','Home','9','10'),
7:00 PM	30	Jan	1999	O002	T01	V016	Home	2	12	('07:00 PM','29','Jan','1999','O002','T01','V016','Home','2','12'),
12:00 PM	13	Feb	1999	O005	T01	V017	Away	14	3	('12:00 PM','12','Feb','1999','O005','T01','V017','Away','14','3'),
3:00 PM	13	Feb	1999	O005	T01	V017	Away	8	3	('03:00 PM','12','Feb','1999','O005','T01','V017','Away','8','3'),
1:00 PM	14	Feb	1999	O005	T01	V017	Away	3	2	('01:00 PM','13','Feb','1999','O005','T01','V017','Away','3','2'),
1:00 PM	21	Feb	1999	O006	T01	V018	Away	4	7	('01:00 PM','20','Feb','1999','O006','T01','V018','Away','4','7'),
2:30 PM	2	Mar	1999	O007	T01	V032	Away	3	3	('02:30 PM','01','Mar','1999','O007','T01','V032','Away','3','3'),
4:00 PM	12	Mar	1999	O010	T01	V019	Away	3	4	('04:00 PM','11','Mar','1999','O010','T01','V019','Away','3','4'),

3. Data Insertion in SQL

- Drop Tables

  These lines of code is created so that in case any other table with a similar name is present we drop them.

```
USE BUDT703_Project_0506_12

--Dropping tables in order opposite to their creation.

DROP TABLE IF EXISTS [Moneyball.Play];
DROP TABLE IF EXISTS [Moneyball.Venue];
DROP TABLE IF EXISTS [Moneyball.Calendar];
DROP TABLE IF EXISTS [Moneyball.Tournament];
DROP TABLE IF EXISTS [Moneyball.Opponent];
```

- Create Tables

```
--Creating table Opponent with opponentId as Primary Key.
CREATE TABLE [Moneyball.Opponent] (
        opponentId CHAR(4) NOT NULL,
        opponentName VARCHAR(50),
        opponentUniversity VARCHAR(99),
        CONSTRAINT pk_Opponent_opponentId PRIMARY KEY (opponentId)
);

 --Creating table Tournament with tournamentId as Primary Key.
CREATE TABLE [Moneyball.Tournament] (
        tournamentId CHAR(3) NOT NULL,
        tournamentName VARCHAR (50),
        CONSTRAINT pk_Tournament_tournamentId PRIMARY KEY (tournamentId)
);

-- Creating table calendar with calendarYear and calendarMonth as Composite Primary Keys.
CREATE TABLE [Moneyball.Calendar](
        calendarMonth CHAR(3) NOT NULL,
        calendarYear DATE NOT NULL,
        CONSTRAINT pk_Calendar_calendarYear_calendarMonth PRIMARY KEY (calendarYear, calendarMonth)
);

 --Creating table Venue with venueId as Primary Key.
CREATE TABLE [Moneyball.Venue](
        venueId CHAR(5) NOT NULL,
        venueName VARCHAR(50),
        venueCity VARCHAR(50),
        venueState CHAR(2),
        CONSTRAINT pk_Venue_venueId PRIMARY KEY (venueId)
);
```

```
 --Creating table of relation Play with calendarYear,calendarMonth,playDate and playTime as composite primary key and
 --opponentId,tournamentId,calendarYear,calendarMonth and venueId as foreign key.
CREATE TABLE [Moneyball.Play](
       playTime CHAR(8) NOT NULL,
       playDate INT NOT NULL,
       calendarMonth CHAR(3) NOT NULL,
       calendarYear DATE NOT NULL,
       opponentId CHAR(4),
       tournamentId CHAR(3),
       venueId CHAR(5),
       playType VARCHAR(10),
       playTerpScore INT,
       playOpponentScore INT,
       CONSTRAINT pk_Play_playTime_playDate_calendarMonth_calendarYear PRIMARY KEY (playTime,playDate,calendarYear,calendarMonth),
       CONSTRAINT fk_Play_opponentId FOREIGN KEY (opponentId)
              REFERENCES [Moneyball.Opponent](opponentId)
              ON DELETE NO ACTION ON UPDATE CASCADE,
       CONSTRAINT fk_Play_tournamentId FOREIGN KEY (tournamentId)
              REFERENCES [Moneyball.Tournament](tournamentId)
              ON DELETE NO ACTION ON UPDATE CASCADE,
       CONSTRAINT fk_Play_calendarYear_calendarMonth FOREIGN KEY (calendarYear, calendarMonth)
              REFERENCES [Moneyball.calendar](calendarYear, calendarMonth)
              ON DELETE NO ACTION ON UPDATE CASCADE,
       CONSTRAINT fk_Play_venueId FOREIGN KEY (venueId)
              REFERENCES [Moneyball.Venue](venueId)
              ON DELETE NO ACTION ON UPDATE CASCADE,
);
```

- Insert Tables.
  These are the lines of code written to insert the values into the respective table of the database.

```
--Inserting Values into the respective tables

INSERT INTO [Moneyball.Opponent](opponentId,opponentName,opponentUniversity) VALUES
('O001','Oklahoma','University of Oklahoma'),
('O002','Jacksonville','Jacksonville University'),
('O003','Auburn','Auburn University'),
('O004','North Carolina A&T','North Carolina A&T University'),
('O005','Elon','Elon University'),


INSERT INTO [Moneyball.Tournament] (tournamentId,tournamentName) VALUES
('T01','Club Friendlies'),
('T02','Big Ten'),
('T03','ACC Championship'),
('T04','NCAA Tournament'),
('T05','NCAA Super Regional'),
```

```
INSERT INTO [Moneyball.Venue](venueId,venueName,venueCity,venueState) VALUES
('V001','USF Baseball Stadium','Tampa','FL'),
('V002','Bob "Turtle" Smith Stadium','College Park','MD'),
('V003','Swayze Field','Oxford','MS'),
('V004','U.S. Bank Stadium','Minneapolis','MN'),
('V005','Bob Hanna Stadium','Newark','DE'),

INSERT INTO [Moneyball.Calendar] (calendarMonth,calendarYear) VALUES
('Jan','2023'),
('Feb','2023'),
('Mar','2023'),
('Apr','2023'),
('May','2023'),
('Jun','2023'),

INSERT INTO [Moneyball.Play]
(playTime,playDate,calendarMonth,calendarYear,opponentId,tournamentId,venueId,playType,playTerpScore,playOpponentScore)
SELECT playTime,playDate,calendarMonth,calendarYear,opponentId,tournamentId,venueId,playType,playTerpScore,playOpponentScore FROM (
VALUES
('03:00 PM','29','Jan','1999','O001','T01','V016','Home','9','10'),
('07:00 PM','30','Jan','1999','O002','T01','V016','Home','2','12'),
('12:00 PM','13','Feb','1999','O005','T01','V017','Away','14','3'),
('03:00 PM','13','Feb','1999','O005','T01','V017','Away','8','3'),
('01:00 PM','14','Feb','1999','O005','T01','V017','Away','3','2'),
('01:00 PM','21','Feb','1999','O006','T01','V018','Away','4','7'),
('02:30 PM','02','Mar','1999','O007','T01','V032','Away','3','3'),
('04:00 PM','12','Mar','1999','O010','T01','V019','Away','3','4'),
('03:00 PM','17','Mar','1999','O011','T01','V002','Home','15','7'),
AS SUB(playTime,playDate,calendarMonth,calendarYear,opponentId,tournamentId,venueId,playType,playTerpScore,playOpponentScore);


--Adding the dervied playResult column

ALTER TABLE [Moneyball.Play] ADD playResult CHAR(1);

UPDATE [Moneyball.Play]
        SET playResult = (CASE
                                        WHEN playTerpScore > playOpponentScore THEN 'W'
                                        WHEN playTerpScore < playOpponentScore THEN 'L'
                                        ELSE 'D'
                                        END)
        FROM [Moneyball.Play]
```

4. SQL Query and Result Interpretation

- **Transaction 1 What is the ratio of game wins based on home, away, and neutral sites?**

  Our query aims to calculate the win ratio for the games based on the location where the game is played, which is the play type: home, neutral or away sites to identify the presence of a home game advantage for the Terps. The query retrieves the win ratio for each play type by counting the number of wins and calculating the ratio over the total number of plays. This is then grouped by the play type and years and ordered by years as well so that the obtained output can be interpreted over the years and trends can be analyzed.

Results Messages

	Year	Play Type	WinRatio
1	1999	Away	0.46
2	1999	Home	0.42
3	2000	Away	0.38
4	2000	Home	0.33
5	2001	Away	0.36
6	2001	Home	0.22
7	2002	Away	0.36
8	2002	Home	0.82
9	2003	Away	0.4
10	2003	Home	0.39
11	2004	Away	0.26
12	2004	Home	0.52
13	2005	Away	0.52
14	2005	Home	0.47
15	2006	Away	0.32
16	2006	Home	0.64
17	2007	Away	0.3
18	2007	Home	0.58
19	2008	Away	0.39
20	2008	Home	0.64

Query executed successfully. doitsqlx.rhsmith.umd.edu,97... AD\ganesaan (115) BUDT703_Project_0506_12 00:00:00 60 rows

- **Transaction 2 What is the team's average margin of wins and losses in the previous years?**

  For this transaction we calculate the average margin of win or loss. First, we take the difference of the scores in the matches where Terps won and divide it by the number of games won. This will give the average win margin. Similarly, we compute the average loss margin with the numerator being the difference of the scores in the matches where Terps lost and divide it by the games terps lost. This is then grouped by year yielding the trend.

	Year	AvgWinMargin	AvgLossMargin
1	1999	6.4	-5.5
2	2000	4.9	-4
3	2001	5.8	-4.7
4	2002	6.6	-5.9
5	2003	4	-6
6	2004	5.1	-5.1
7	2005	4.8	-4.2
8	2006	2.9	-5
9	2007	5.9	-5.4
10	2008	5	-5.2
11	2009	4.9	-5.5
12	2010	3.8	-6.7
13	2011	4.3	-5.1
14	2012	4.4	-3.3
15	2013	3.8	-3.5
16	2014	4.7	-4.9
17	2015	4.4	-2.8
18	2016	3.7	-3.1
19	2017	4.7	-4.1
20	2018	4.9	-4.7

Query executed successfully. doitsqlx.rhsmith.umd.edu,97... AD\ganesaan (115) BUDT703_Project_0506_12 00:00:00 25 rows

- **Transaction 3 What is the ratio of game wins in Big Ten conference games?**

  In this transaction we aim to see the superiority of the Terps baseball team at the Big Ten Conference. We start of by calculating the number of wins and divide it by the total number of matches played in the tournament. Subsequently we filter out the results for the BIG Ten tournament and then group by tournament name and year.

	Year	WinRatio
1	1999	0.33
2	2006	0.27
3	2007	0.24
4	2008	0.3
5	2009	0.33
6	2010	0.21
7	2011	0.17
8	2012	0.33
9	2013	0.37
10	2015	0.58
11	2016	0.54
12	2017	0.63
13	2018	0.39
14	2019	0.5
15	2021	0.64
16	2022	0.78
17	2023	0.71

Query executed successfully. doitsqlx.rhsmith.umd.edu,97... AD\ganesaan (115) BUDT703_Project_0506_12 00:00:00 17 rows

- **Transaction 4 What is the team's Winning rate against opponents where atleast 10 games being played?**

  In order to analyze the superiority of the terps team against opponents we face regularly (matches played overall >=10) we create a query to obtain the win margin. We start of by calculating the number of wins and divide it by the total number of matches played against that opponent using the data from the Play and Opponent table. We set a condition where the number of matches played is greater than or equal to 10 and we then sort by the descending order of win ratio.

	Opponent Name	WinRatio
1	Coppin State	1
2	Maryland-Eastern Shore	1
3	Navy	1
4	Princeton	1
5	UMBC	0.92
6	Bryant	0.87
7	VCU	0.76
8	Purdue	0.74
9	Penn State	0.73
10	Georgetown	0.73
11	George Washington	0.73
12	Ohio State	0.71
13	Rutgers	0.68
14	Old Dominion	0.67
15	Minnesota	0.67
16	Delaware	0.65
17	William & Mary	0.64
18	Towson	0.63
19	Michigan State	0.63
20	Northwestern	0.62

Query executed successfully. doitsqlx.rhsmith.umd.edu,97... AD\ganesaan (115) BUDT703_Project_0506_12 00:00:00 43 rows

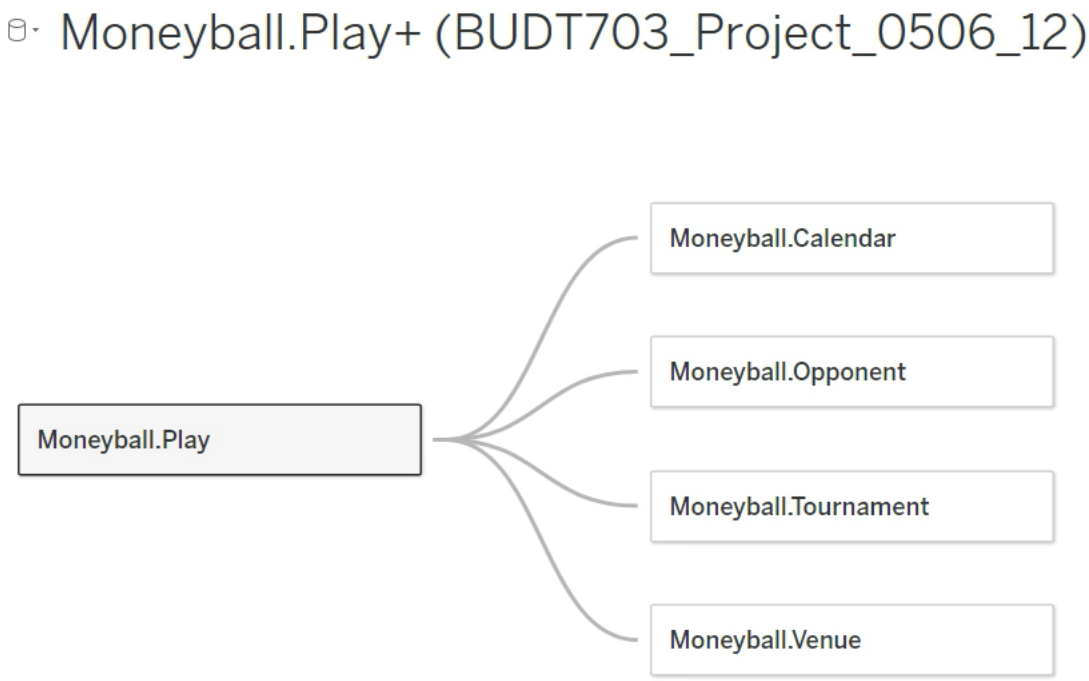
- **Transaction 5 What is the win ratio over the years across different States?**

  In order to analyze the Identify the states where our team traditionally performed well versus, we create a query to obtain the win margin. We start of by calculating the number of wins and divide it by the total number of matches in a particular state using the data from the Play and Venue table. We group the data by the state and then sort it by the descending order of win ratio.

	State	WinRatio
1	NY	1
2	DC	0.86
3	NJ	0.83
4	PA	0.67
5	CA	0.64
6	IN	0.63
7	MD	0.61
8	IA	0.6
9	NE	0.57
10	IL	0.56
11	DE	0.54
12	AL	0.5
13	MI	0.5
14	OH	0.5
15	TN	0.5
16	TX	0.5
17	MN	0.44
18	SC	0.43
19	NC	0.42
20	VA	0.38

Query executed successfully. doitsqlx.rhsmith.umd.edu,97... AD\ganesaan (115) BUDT703_Project_0506_12 00:00:00 27 rows

5. Creating Tableau Views

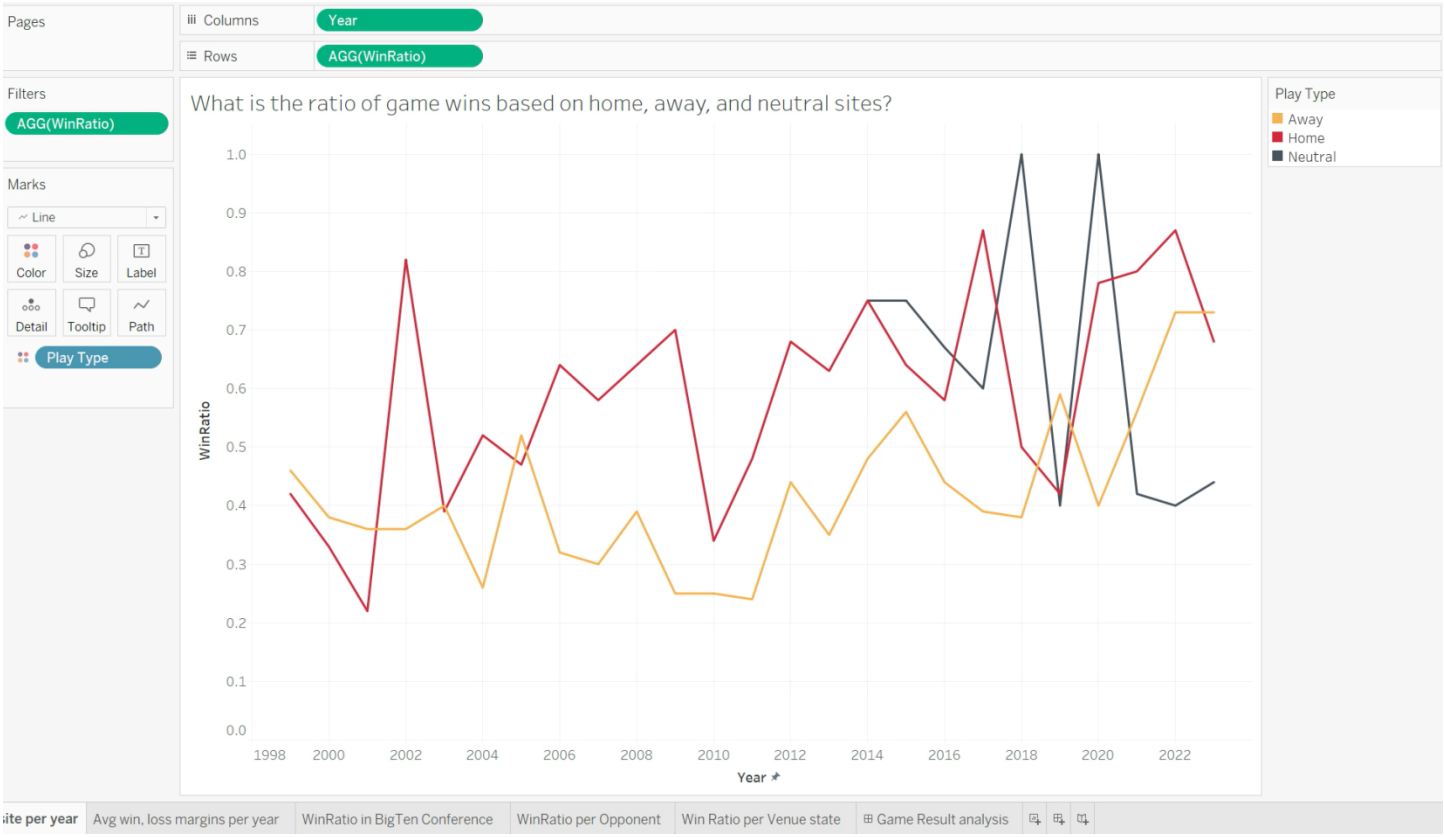


We first connect the database to the tableau server and then make the connections as shown above, which imitates the quaternary relation we have among the tables.

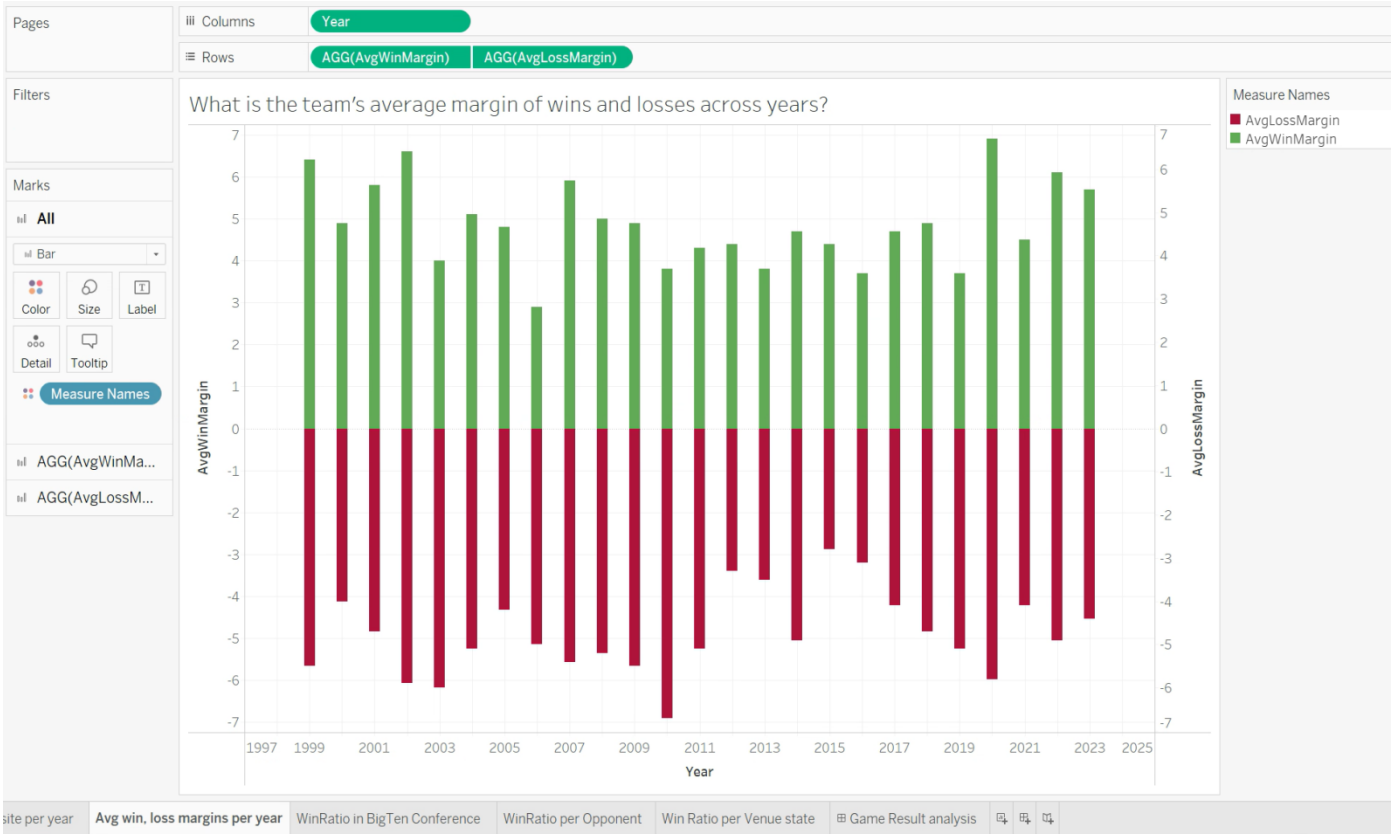
- Create calculated fields for usage in Tableau views.
  - Games = Count([PlayResult])
  - AvgLossMargin - AVG(IF [Play Result]= 'L' THEN [Play Terp Score] - [Play Opponent Score] ELSE 0 END)
  - AvgWinMargin - AVG(IF [Play Result]= 'W' THEN [Play Terp Score]-[Play Opponent Score] ELSE 0 END)

- WinRatio - ROUND(SUM(IF [Play Result]= 'W' THEN 1 ELSE 0 END)/COUNT([Play Result]),2) Wins - SUM(IF [Play Result]= 'W' THEN 1 ELSE 0 END

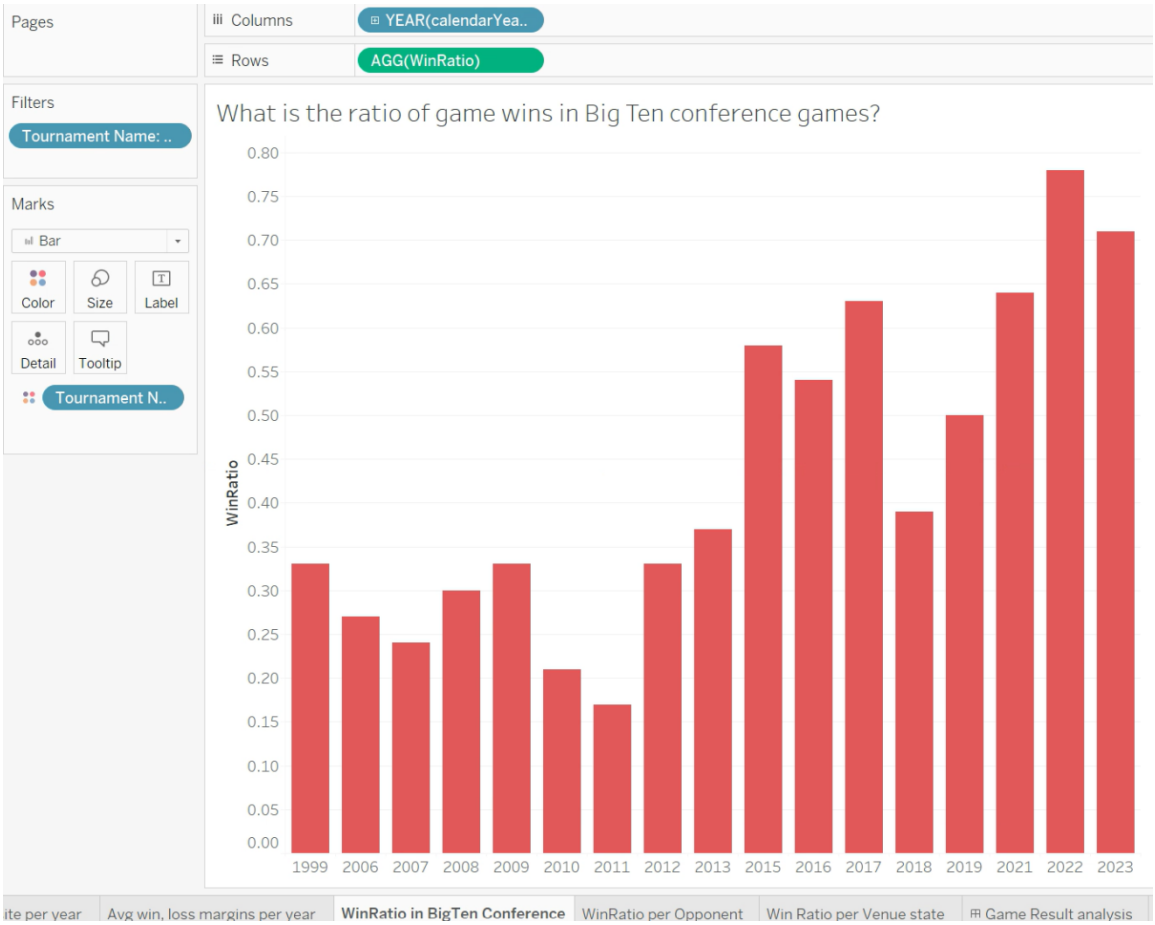
- Tableau Visualizations
- **Transaction 1 What is the ratio of game wins based on home, away, and neutral sites?**



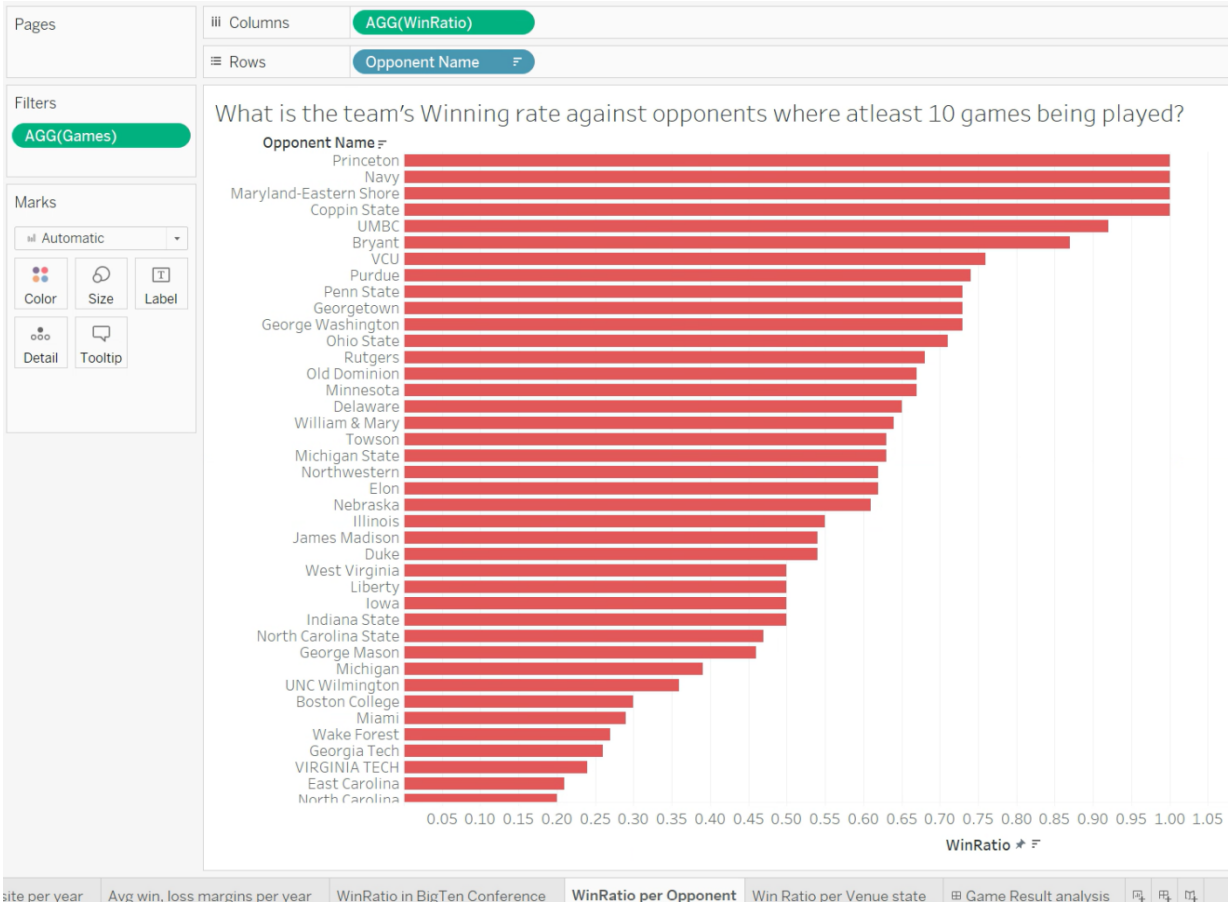
- **Transaction 2 What is the team's average margin of wins and losses in the previous years?**



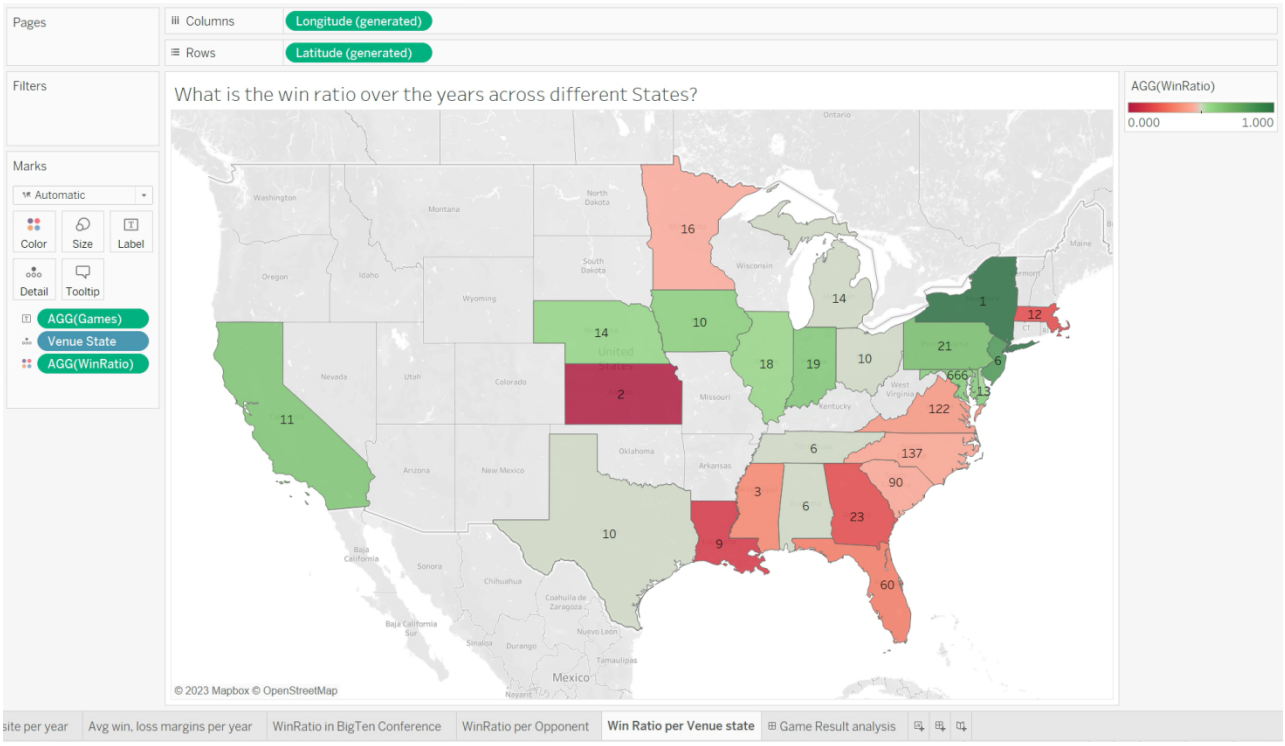
- **Transaction 3 What is the ratio of game wins in Big Ten conference games?**



- **Transaction 4 What is the team's Winning rate against opponents where atleast 10 games being played?**

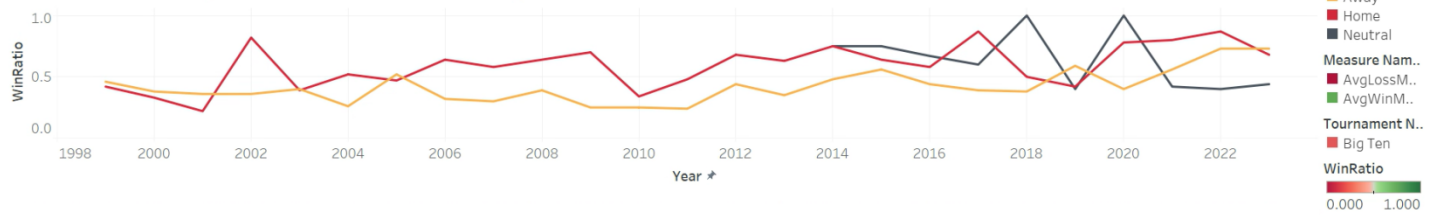


- **Transaction 5 What is the win ratio over the years across different States?**
  In the Data pane, click the data type icon next to 'Venue State' field under 'Moneryball.Venue', select Geographic Role, and then select the 'State/Province' role to assign to the field.

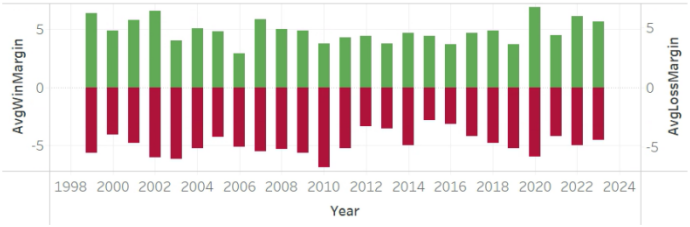


- **Tableau Dashboard**

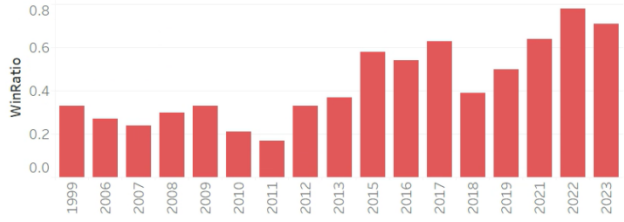
What is the ratio of game wins based on home, away, and neutral sites?



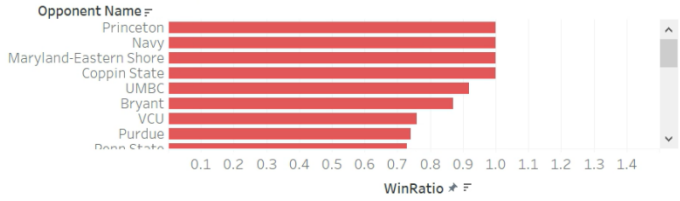
What is the team's average margin of wins and losses across years?



What is the ratio of game wins in Big Ten conference games?



What is the team's Winning rate against opponents where atleast 10 games being played?



What is the win ratio over the years across different States?