Data Quality Assessment:

Dataset: NFL Scores and Betting Data (nfl_teams.csv and spreadspoke_scores.csv)
Source: Kaggle.com, compiled from ESPN, NFL.com, Pro-Football-Reference, NOAA, NFLweather.com, and other betting archives.

1. Data Quality Assessment and Improvement

Accuracy:
Data is pulled from reputable sources and cross-referenced (e.g., betting lines from multiple websites; weather from NOAA and NFLweather.com).

Completeness:
The dataset includes NFL scores from 1966 and betting odds since 1979. However, `team_division` is missing in some `nfl_teams.csv` rows.

Consistency:
Several fields require standardization:

- `schedule_date` is a string, not a date.
- `schedule_week` is 96% numeric, but 4% are labels (e.g., “Superbowl,” “Wildcard”).
- `over_under_line` includes 0.4% invalid values (e.g., space characters).
- `weather_detail` contains inconsistent formatting; 0.8% use pipe delimiters (e.g., "Rain | Fog").

Timeliness:
Data is updated weekly, with a maximum potential 7-day lag post-game.

Uniqueness:
No duplicate rows were found.

Summary:
The dataset is accurate and timely but has minor completeness gaps and several consistency issues involving date, weather, and number formatting.

Transformations / Cleaning Applied:

- In `nfl_teams.csv`, missing team division values were labeled “not applicable” for now.
- `over_under_line` blank spaces were cleaned and converted to numeric.
- `weather_detail` was split by delimiter into separate boolean columns (e.g., Rain, Fog).

---

2. Database Schema Design (SQL DDL)

While all of the completeness issues were resolvable by the actions taken in stage 1, to resolve lingering consistency issues, I designed a star schema during ETL, splitting the data into one fact table (`Scores`) and three dimension tables: `Teams`, `Stadia`, and `Weather`. See fig 1.

- `Teams` and `Stadia` are Type 2 dimensions (include `is_current`, `valid_from`, `valid_to`) to handle changing values.
- `Weather` is a static reference table, as weather types are fixed. New types may be added, but existing types do not have any variables that might change over time.

Date Handling:
Dates stored as `varchar(8)` in ISO 112 format to prevent regional conflicts during export (e.g., Excel misreading US/UK date formats). This remediates data consistency issues with the date fields.

Rationale:
This schema improves performance (normalisation, more efficient datatypes), data quality (specifically with regards to consistency, as only one column remains with data consistency issues - *`schedule_week`*), and maintainability. SQL DDL was written and tested using MS SQL Server. Field types and definitions documented below, with sample values included.

Further Improvements:

To turn this into a proper ETL pipeline, the views created would populate staging tables and then a merge usp used to check for differences between staging and production tables, updating each where changes are identified and updating iscurrent and validfrom/validto for deprecated rows. This would allow post and pre-match changes to the tables to be reflected accurately.

`schedule_week` could be spun out into a schedule dimension table where dates have a match_type dimension and the Score fact table could link to it via a 1:1 PK/SK relationship to resolve the last consistency issue.

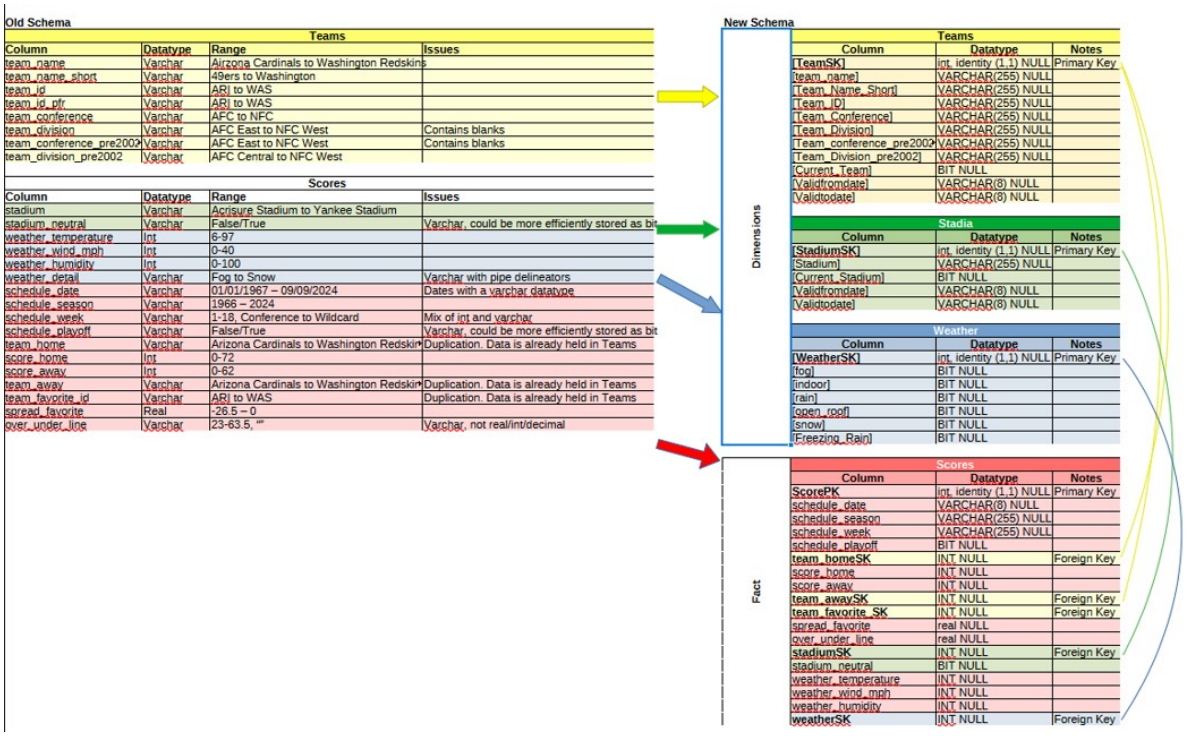


Figure 1: Old Schema to New Schema Overview

Screenshots and SQL Samples:
This submission includes before/after dataset screenshots, schema diagrams, and annotated SQL code showing table creation, data types, and field-level assumptions.

```sql
CREATE TABLE [Teams]
(
[TeamSK] int identity (1,1),
[team_name] VARCHAR(255) NULL,
[Team_Name_Short] VARCHAR(255) NULL,
[Team_ID] VARCHAR(255) NULL,
[Team_Conference] VARCHAR(255) NULL,
[Team_Division] VARCHAR(255) NULL,
[Team_conference_pre2002] VARCHAR(255) NULL,
[Team_Division_pre2002] VARCHAR(255) NULL,
[Current_Team] BIT NULL,
[Validfromdate] VARCHAR(8) NULL,
[Validtodate] VARCHAR(8) NULL
);

CREATE VIEW vpopulate_Teams AS
SELECT
team_name AS [Team_Name],
team_name_short AS [Team_Name_Short],
team_id AS [Team_ID],
team_conference AS [Team_Conference],
team_division AS [Team_Division],
team_conference AS [Team_conference_pre2002],
team_division_pre2002 AS [Team_Division_pre2002],
CASE When team_division = 'Not Applicable' Then 0 Else 1 End AS [Current_Team],
19000101 AS [Validfromdate],
29991231 AS [Validtodate]

FROM nfl_teams_amended;

INSERT INTO Teams
SELECT *
FROM vpopulate_Teams;

-- Manual updates for no longer current team names

UPDATE teams SET validtodate = 31081996 WHERE team_name = 'Houston Oilers';
UPDATE teams SET validfromdate = 01091996, validtodate = 31081998 WHERE team_name = 'Tennessee Oilers';
UPDATE teams SET validfromdate = 01091998 WHERE team_name = 'Tennessee Titans';

-- Boston / New England Patriots
UPDATE teams SET validtodate = 31081970 WHERE team_name = 'Boston Patriots';
UPDATE teams SET validfromdate = 01091971 WHERE team_name = 'New England Patriots';

-- St. Louis / Phoenix / Arizona Cardinals
UPDATE teams SET validtodate = 31081987 WHERE team_name = 'St. Louis Cardinals';
UPDATE teams SET validfromdate = 01091988, validtodate = 31081993 WHERE team_name = 'Phoenix Cardinals';
UPDATE teams SET validfromdate = 01091994 WHERE team_name = 'Arizona Cardinals';

-- Baltimore / Indianapolis Colts
UPDATE teams SET validtodate = 31081983 WHERE team_name = 'Baltimore Colts';
UPDATE teams SET validfromdate = 01091984 WHERE team_name = 'Indianapolis Colts';

-- Oakland / Los Angeles / Las Vegas Raiders
UPDATE teams SET validtodate = 31081981 WHERE team_name = 'Oakland Raiders';
UPDATE teams SET validfromdate = 01091982, validtodate = 31081994 WHERE team_name = 'Los Angeles Raiders';
UPDATE teams SET validfromdate = 01091995, validtodate = 31082019 WHERE team_name = 'Oakland Raiders';
UPDATE teams SET validfromdate = 01092020 WHERE team_name = 'Las Vegas Raiders';
```

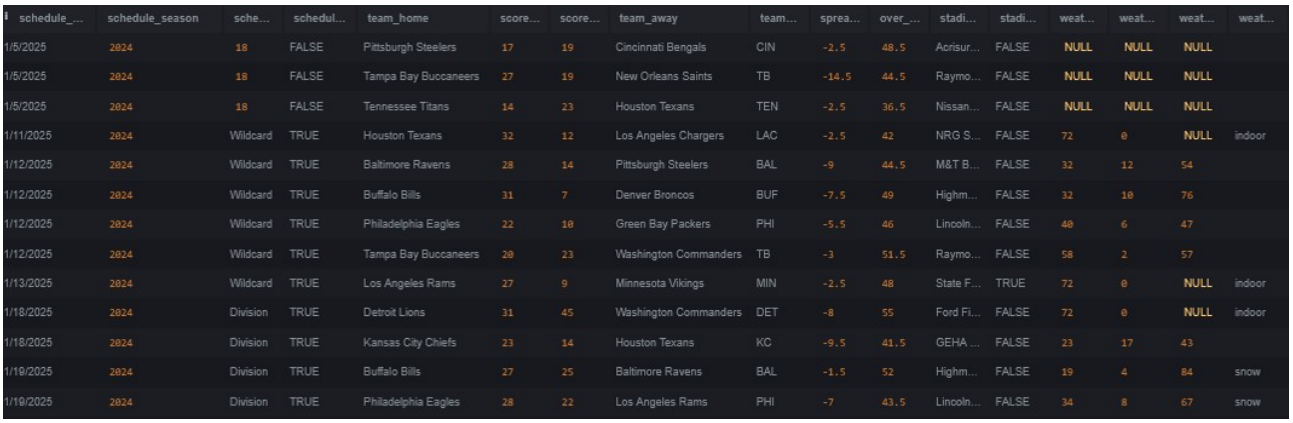
Figure 2: Teams Table and View Create and populate code

```sql
-- Create Stadia table
CREATE Table Stadia
(
StadiumSK int identity (1,1),
Stadium Varchar(255) null,
Current_stadium bit null,
validfromdate varchar(8),
validtodate varchar(8),
);

-- create view to populate it
CREATE VIEW vpopulate_Stadia
AS
SELECT DISTINCT stadium,
1 AS current_stadium,
19000101 AS validfromdate,
29991231 AS validtoDate
FROM spreadspoke_scores_amended;

-- Populate it
INSERT INTO Stadia
SELECT * FROM vpopulate_Stadia;


-- create a weather reference table
CREATE TABLE [Weather]
(
[WeatherSK] int identity (1,1),
[fog] bit null,
[indoor] bit null,
[rain] bit null,
[open_roof] bit null,
[snow] bit NULL,
[freezing_rain] bit NULL
)

-- create view to populate it with every combination of weather
CREATE VIEW vpopulate_Weather
AS
SELECT a.bit AS fog,
       b.bit AS indoor,
       c.bit AS rain,
       d.bit AS open_roof,
       e.bit AS snow,
       f.bit AS freezing_rain
FROM (SELECT 0 AS bit UNION ALL SELECT 1) a
CROSS JOIN (SELECT 0 AS bit UNION ALL SELECT 1) b
CROSS JOIN (SELECT 0 AS bit UNION ALL SELECT 1) c
CROSS JOIN (SELECT 0 AS bit UNION ALL SELECT 1) d
CROSS JOIN (SELECT 0 AS bit UNION ALL SELECT 1) e
CROSS JOIN (SELECT 0 AS bit UNION ALL SELECT 1) f

-- Populate it
INSERT INTO Weather
SELECT * FROM vpopulate_Weather
```

Figure 3: Weather and Stadia Table and View SQL commands (including a cartesian join for weather)

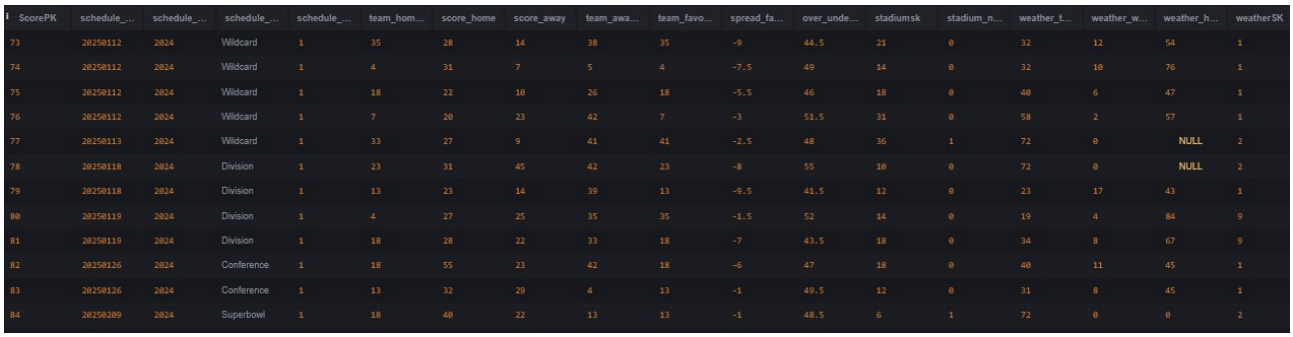
```sql
CREATE TABLE [Scores]
(
[ScorePK] int identity (1,1),
schedule_date varchar(8) null,
schedule_season varchar(4) null,
schedule_week Varchar(255) null,
schedule_playoff bit null,
team_homeSK int null,
  score_home int null,
  score_away int null,
  team_awaySK int null,
  team_favoriteSK int null,
  spread_favorite real null,
  over_under_line real null,
  stadiumsk int null,
  stadium_neutral bit null,
  weather_temperature int null,
  weather_wind_mph int null,
  weather_humidity int null,
  weatherSK int null)

CREATE VIEW vpopulate_Scores AS

SELECT
Convert (Varchar(8), Cast (s.schedule_date AS Date),112) AS schedule_date,
s.schedule_season AS schedule_season,
s.schedule_week AS schedule_week,
CASE WHEN s.schedule_playoff = 'FALSE' then 0 ELSE 1 END AS schedule_playoff,
ht.TeamSK AS team_homesk,
s.score_home AS score_home,
s.score_away AS score_away,
at.TeamSK AS team_awaysk,
ft.TeamSK AS team_favoritesk,
s.spread_favorite AS spread_favoriteSK,
s.over_under_line AS over_under_line,
st.StadiumSK AS stadiumSK,
CASE WHEN s.stadium_neutral = 'FALSE' then 0 ELSE 1 END AS stadium_neutral,
s.weather_temperature AS weather_temperature,
s.weather_wind_mph AS weather_wind_mph,
s.weather_humidity AS weather_humidity,
w.WeatherSK AS weathersk

 FROM spreadspoke_scores_amended S
 LEFT JOIN teams ht ON s.team_home = ht.team_name
         AND CONVERT(Varchar(8), Cast (s.schedule_date AS Date),112) >= ht.Validfromdate
         AND convert(Varchar(8), Cast (s.schedule_date AS Date),112) < ht.Validtodate
 LEFT JOIN teams at ON s.team_away = at.team_name
         AND CONVERT(Varchar(8), Cast (s.schedule_date AS Date),112) >= at.Validfromdate
         AND convert(Varchar(8), Cast (s.schedule_date AS Date),112) < at.Validtodate
  LEFT JOIN teams ft ON s.team_favorite_id = ft.team_id
         AND CONVERT(Varchar(8), Cast (s.schedule_date AS Date),112) >= ft.Validfromdate
         AND convert(Varchar(8), Cast (s.schedule_date AS Date),112) < ft.Validtodate
 LEFT JOIN Weather w ON ISNULL(s.weather_Fog, 0) = ISNULL(w.fog, 0)
   AND ISNULL(s.weather_Indoor, 0) = ISNULL(w.indoor, 0)
   AND ISNULL(s.weather_Rain, 0) = ISNULL(w.rain, 0)
   AND ISNULL(s.open_Roof, 0) = ISNULL(w.open_roof, 0)
   AND ISNULL(s.weather_Snow, 0) = ISNULL(w.snow, 0)
   AND ISNULL(s.weather_FreezingRain, 0) = ISNULL(w.freezing_rain, 0)
Left Join Stadia st ON s.stadium = st.stadium
  AND CONVERT(Varchar(8), CAST(s.schedule_date AS date),112) between st.validfromdate and st.validtodate
```

Figure 4: Scores Fact Table and View Create



schedule_...	schedule_season	sche...	schedul...	team_home	score...	score...	team_away	team...	sprea...	over_...	stadi...	stadi...	weat...	weat...	weat...	weat...
1/5/2025	2024	18	FALSE	Pittsburgh Steelers	17	19	Cincinnati Bengals	CIN	-2.5	48.5	Acrisur...	FALSE	NULL	NULL	NULL	
1/5/2025	2024	18	FALSE	Tampa Bay Buccaneers	27	19	New Orleans Saints	TB	-14.5	44.5	Raymo...	FALSE	NULL	NULL	NULL	
1/5/2025	2024	18	FALSE	Tennessee Titans	14	23	Houston Texans	TEN	-2.5	36.5	Nissan...	FALSE	NULL	NULL	NULL	
1/11/2025	2024	Wildcard	TRUE	Houston Texans	32	12	Los Angeles Chargers	LAC	-2.5	42	NRG S...	FALSE	72	0	NULL	indoor
1/12/2025	2024	Wildcard	TRUE	Baltimore Ravens	28	14	Pittsburgh Steelers	BAL	-9	44.5	M&T B...	FALSE	32	12	54	
1/12/2025	2024	Wildcard	TRUE	Buffalo Bills	31	7	Denver Broncos	BUF	-7.5	49	Highm...	FALSE	32	10	76	
1/12/2025	2024	Wildcard	TRUE	Philadelphia Eagles	22	10	Green Bay Packers	PHI	-5.5	46	Lincoln...	FALSE	40	6	47	
1/12/2025	2024	Wildcard	TRUE	Tampa Bay Buccaneers	20	23	Washington Commanders	TB	-3	51.5	Raymo...	FALSE	58	2	57	
1/13/2025	2024	Wildcard	TRUE	Los Angeles Rams	27	9	Minnesota Vikings	MIN	-2.5	48	State F...	TRUE	72	0	NULL	indoor
1/18/2025	2024	Division	TRUE	Detroit Lions	31	45	Washington Commanders	DET	-8	55	Ford Fi...	FALSE	72	0	NULL	indoor
1/18/2025	2024	Division	TRUE	Kansas City Chiefs	23	14	Houston Texans	KC	-9.5	41.5	GEHA ...	FALSE	23	17	43	
1/19/2025	2024	Division	TRUE	Buffalo Bills	27	25	Baltimore Ravens	BAL	-1.5	52	Highm...	FALSE	19	4	84	snow
1/19/2025	2024	Division	TRUE	Philadelphia Eagles	28	22	Los Angeles Rams	PHI	-7	43.5	Lincoln...	FALSE	34	8	67	snow

Figure 5: Old Scores Table



ScorePK	schedule_...	schedule_...	schedule_...	schedule_...	team_hom...	score_home	score_away	team_awa...	team_favo...	spread_fa...	over_unde...	stadiumsk	stadium_n...	weather_t...	weather_w...	weather_h...	weatherSK
73	20250112	2024	Wildcard	1	35	28	14	38	35	-9	44.5	21	0	32	12	54	1
74	20250112	2024	Wildcard	1	4	31	7	5	4	-7.5	49	14	0	32	10	76	1
75	20250112	2024	Wildcard	1	18	22	10	26	18	-5.5	46	18	0	40	6	47	1
76	20250112	2024	Wildcard	1	7	20	23	42	7	-3	51.5	31	0	58	2	57	1
77	20250113	2024	Wildcard	1	33	27	9	41	41	-2.5	48	36	1	72	0	NULL	2
78	20250118	2024	Division	1	23	31	45	42	23	-8	55	10	0	72	0	NULL	2
79	20250118	2024	Division	1	13	23	14	39	13	-9.5	41.5	12	0	23	17	43	1
80	20250119	2024	Division	1	4	27	25	35	35	-1.5	52	14	0	19	4	84	9
81	20250119	2024	Division	1	18	28	22	33	18	-7	43.5	18	0	34	8	67	9
82	20250126	2024	Conference	1	18	55	23	42	18	-6	47	18	0	40	11	45	1
83	20250126	2024	Conference	1	13	32	29	4	13	-1	49.5	12	0	31	8	45	1
84	20250209	2024	Superbowl	1	18	40	22	13	13	-1	48.5	6	1	72	0	0	2

Figure 6: New Scores Table