Data Premier League
Data Visualization and Prediction

Scenario: Formula 1 Driver Performance Prediction Challenge

Objective:

The roar of engines fills the air as the Formula 1 circus arrives at a legendary circuit. In the paddock, the team principal strides into the analytics room, clutching a stack of race data spanning seven decades, from 1950 to 2024. With an urgent tone, they demand:

"Predict the finishing positions of Formula 1 drivers in an upcoming race using historical data and advanced analytics. The sponsors need clarity, and our strategy depends on it. Can you deliver?"

This challenge isn't just about crunching numbers; it's about leveraging the power of data storytelling, statistical analysis, and machine learning. Armed with a comprehensive dataset—detailing driver performances, team histories, pit stop strategies,etc… —you and your analytics team must uncover the secrets of race outcomes.

Tasks:

1. Data Exploration and Preprocessing

- Analyze dataset structure and trends from 1950-2024.
- Handle missing or inconsistent data.
- Conduct statistical tests.

2. Feature Engineering

- Create new predictive features such as:
    - **Driver Consistency** – Average finishing position & qualifying performance.

    - **Team Strength** – Constructor points & reliability trends.
    - **Track Complexity** – Circuit overtaking statistics.

3. Model Development & Evaluation

- Build machine learning models to predict race outcomes.
- Train and validate models using historical data.
- Evaluate using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R^2 Score and Cross-Validation Techniques.

4. Insights & Visualization

- Identify key race outcome influencers:
    - **Constructor reliability & team dynamics.**
    - **Driver adaptability to circuits & weather.**
    - **Strategic pit stop decisions.**
- Create impactful visualizations:
    - **Heatmaps** – Performance correlations.
    - **Historical dominance charts** – Constructors & drivers.
    - **Scatterplots** – Comparing team budgets & race results.

5. Presentation & Documentation

- Summarize findings with engaging visualizations.
- Deliver a polished presentation of model performance & insights.

Dataset Structure:

The dataset consists of all information on the Formula 1 World Championship from 1950 till the latest 2024 season. The dataset is organized into 14 csv files. Each csv file contains attributes as follows:

circuits.csv

- **circuitId:** A unique identifier for each circuit.
- **circuitRef:** A reference name or short code for the circuit.
- **name:** The full name of the circuit.
- **location:** The city or locality where the circuit is situated.
- **country:** The country where the circuit is located.
- **lat:** Latitude coordinate of the circuit.
- **lng:** Longitude coordinate of the circuit.
- **alt:** Altitude (in meters) of the circuit above sea level.

- **url:** A web link for more information about the circuit.

constructor_results.csv

- **constructorResultsId**: A unique identifier for each constructor result.
- **raceId**: A unique identifier linking the result to a specific race.
- **constructorId**: A unique identifier linking the result to a specific constructor/team.
- **points**: The number of points earned by the constructor in the specific race.
- **status**: The status of the constructor's participation or result in the race, where\N is NULL and ‘D’ denotes disqualified.

constructor_standings.csv

- **constructorStandingsId**: A unique identifier for each constructor's standings entry.
- **raceId**: A unique identifier linking the standings to a specific race.
- **constructorId**: A unique identifier for the constructor/team.
- **points**: The total number of points the constructor has accumulated up to this race.
- **position**: The constructor's numerical rank in the standings (e.g., 1 for first place).
- **positionText**: A textual representation of the position, which may provide additional details (e.g., "1st").
- **wins**: The number of race wins the constructor has achieved up to this race.

constructors.csv

- **constructorId:** A unique identifier for each constructor/team.
- **constructorRef:** A reference or short code for the constructor (likely a simplified or shorthand version of the name).
- **name:** The full name of the constructor/team.
- **nationality:** The nationality of the constructor (e.g., British, Italian, etc.).
- **url:** ***A web link providing more information about the constructor.***

driver_standings.csv

- **driverStandingsId**: A unique identifier for each driver's standings entry.
- **raceId**: A unique identifier linking the standings to a specific race.
- **driverId**: A unique identifier for the driver.
- **points**: The total number of points the driver has accumulated up to this race.
- **position**: The driver's numerical rank in the standings (e.g., 1 for first place).
- **positionText**: A textual representation of the position, potentially including additional details (e.g., "1st" or "N/A").
- **wins**: The number of race wins the driver has achieved up to this race.

drivers.csv

- **driverId**: A unique identifier for each driver.
- **driverRef**: A reference or short code for the driver (likely a simplified or shorthand version of their name).
- **number**: The racing number associated with the driver (if applicable).
- **code**: A three-letter abbreviation representing the driver (commonly used in racing).
- **forename**: The driver's first name.
- **surname**: The driver's last name.
- **dob**: The driver's date of birth.
- **nationality**: The driver's nationality.
- **url**: A web link providing more information about the driver.

lap_times.csv

- **raceId**: A unique identifier linking the data to a specific race.
- **driverId**: A unique identifier for the driver.
- **lap**: The lap number (e.g., 1, 2, 3, etc.).
- **position**: The driver's position at the end of the specific lap.
- **time**: The recorded lap time in a human-readable format (e.g., `1:22.345`).
- **milliseconds**: The recorded lap time in milliseconds for precise numerical analysis.

pit_stops.csv

- **raceId**: A unique identifier linking the data to a specific race.
- **driverId**: A unique identifier for the driver making the pit stop.
- **stop**: The number of the pit stop for the driver during the race (e.g., 1 for the first stop, 2 for the second stop, etc.).
- **lap**: The lap on which the pit stop occurred.
- **time**: The time of day or race time when the pit stop happened (e.g., `14:23:45` or `1:12:34`).
- **duration**: The total time taken for the pit stop (e.g., `22.345` seconds).
- **milliseconds**: The duration of the pit stop in milliseconds for precise analysis.

qualifying.csv

- **qualifyId**: A unique identifier for the qualifying session entry.
- **raceId**: A unique identifier linking the data to a specific race.
- **driverId**: A unique identifier for the driver.
- **constructorId**: A unique identifier for the constructor/team.
- **number**: The driver's racing number.
- **position**: The driver's qualifying position (e.g., 1 for pole position, 2 for second place, etc.).

- **q1**: The driver's recorded lap time in the first qualifying session (Q1).
- **q2**: The driver's recorded lap time in the second qualifying session (Q2), if applicable.
- **q3**: The driver's recorded lap time in the third qualifying session (Q3), if applicable.

races.csv

- **raceId**: A unique identifier for each race.
- **year**: The year in which the race took place.
- **round**: The round number of the race within the season (e.g., 1 for the first race of the season).
- **circuitId**: A unique identifier linking the race to a specific circuit.
- **name**: The official name of the race (e.g., "Monaco Grand Prix").
- **date**: The date of the race.
- **time**: The start time of the race (if applicable).
- **url**: A link to more information about the race.
- **fp1_date**: The date of the first practice session (FP1).
- **fp1_time**: The time of the first practice session (FP1).
- **fp2_date**: The date of the second practice session (FP2).
- **fp2_time**: The time of the second practice session (FP2).
- **fp3_date**: The date of the third practice session (FP3).
- **fp3_time**: The time of the third practice session (FP3).
- **quali_date**: The date of the qualifying session.
- **quali_time**: The time of the qualifying session.
- **sprint_date**: The date of the sprint race (if applicable).
- **sprint_time**: The time of the sprint race (if applicable).

results.csv

- **resultId**: A unique identifier for each result entry.
- **raceId**: A unique identifier linking the result to a specific race.
- **driverId**: A unique identifier for the driver.
- **constructorId**: A unique identifier for the constructor/team.
- **number**: The racing number of the driver.
- **grid**: The starting position (grid) of the driver in the race.
- **position**: The final position of the driver in the race (e.g., 1st, 2nd, 3rd).
- **positionText**: A textual representation of the position (e.g., "1st", "2nd").
- **positionOrder**: A numerical order of positions (useful for sorting or ranking).
- **points**: The number of points earned by the driver in the race.
- **laps**: The total number of laps completed by the driver in the race.
- **time**: The total time taken by the driver to finish the race (e.g., `1:30:45`).
- **milliseconds**: The total race time in milliseconds for more precise analysis.

- **fastestLap**: The lap number during which the driver recorded their fastest lap.
- **rank**: The rank of the driver based on the fastest lap time.
- **fastestLapTime**: The time taken by the driver for their fastest lap.
- **fastestLapSpeed**: The speed of the driver during their fastest lap (e.g., in km/h or mph).
- **statusId**: A code representing the driver's race status (e.g., finished, retired, disqualified).

seasons.csv

- **year**: The year of the racing season or event.
- **url**: A web link providing more information about the race season or event for that particular year.

sprint_results.csv

- **resultId**: A unique identifier for each result entry.
- **raceId**: A unique identifier linking the result to a specific race.
- **driverId**: A unique identifier for the driver.
- **constructorId**: A unique identifier for the constructor/team.
- **number**: The racing number of the driver.
- **grid**: The starting position (grid) of the driver in the race.
- **position**: The final position of the driver in the race (e.g., 1st, 2nd).
- **positionText**: A textual representation of the position (e.g., "1st", "2nd").
- **positionOrder**: The numerical order of positions for sorting or ranking purposes.
- **points**: The number of points awarded to the driver based on their finishing position in the race.
- **laps**: The total number of laps completed by the driver.
- **time**: The total time taken by the driver to finish the race (e.g., `1:30:45`).
- **milliseconds**: The total time in milliseconds for more precise data.
- **fastestLap**: The lap number during which the driver achieved their fastest lap time.
- **fastestLapTime**: The time taken for the fastest lap.
- **statusId**: A code representing the race status of the driver (e.g., "Finished," "Retired").

status.csv

- **statusId**: A unique identifier for each status code (often numeric or alphanumeric).
- **status**: A textual description of the status, such as "Finished," "Retired," "Disqualified," or other race-related outcomes.

Key Insights to Explore

1. **Constructor Dominance Trends** – Team performance over decades.
2. **Pit Stop Strategy** – Correlation between pit stops and race outcomes.

3. **Driver Consistency** – Who performs consistently across seasons?
4. **Circuit Complexity Analysis** – Which tracks favor specific drivers/teams?
5. **Predicting Championship Contenders** – Identifying title favorites.

Problem Statements:

Participants are invited to employ their imagination, prediction skills, and visualization techniques with the given dataset. As you explore the intricacies of the Formula 1 data through the questions provided, envision and predict compelling insights. Let your creativity and analytical prowess shine as you navigate through the diverse dimensions of the Racing world encapsulated in this dataset.

1. **Driver & Constructor Performance**
    - Identify dominant drivers and constructors by analyzing win ratios and podium finishes.
    - Assess the relationship between career longevity and success metrics (wins, podiums, points).
2. **Qualifying vs. Race Performance**
    - How does starting grid position impact final race results? Do certain drivers excel at making up positions?
3. **Pit Stop Strategies**
    - Evaluate optimal pit stop frequency and timing for race success.
    - Analyze pit stop efficiency and its influence on race outcomes.
4. **Head-to-Head Driver Analysis:**
    - Which rivalries have been the most competitive? Identify head-to-head stats based on race finishes.

5. **Hypothetical Driver Swaps:**
    - Swap two drivers between different teams and predict the impact on team and driver standings.
6. **Driver Movements & Team Networks:**
    - Map driver transitions across teams use network graph for visualizations.
7. **Team Performance Comparison:**
    - Compare team success rates against different opponents with and without considering circuit factor.
8. **Driver Consistency in Race Performance:**
    - Identify drivers with consistent top finishes and those with fluctuating results.
9. **Lap Time Efficiency:**
    - Compare lap times across different circuits and identify which teams maximize efficiency.
10. **Best Team Lineup:**
    - Build the best possible team lineup based on driver performance trends.
11. **Predictions for 2025 Season:**
    - Who will win the Drivers' and Constructors' Championship based on historical and current data?
12. **Struggling Teams Analysis:**
    - Predict which team is most likely to underperform in the upcoming 2025 season based on historical trends.
13. **Driver-Specific Track Struggles:**
    - Identify circuits where specific drivers consistently struggle or excel.

14. **Championship Retention Probability:**
    - What is the probability that this season's winner will retain the title in the next season? Analyze historical trends of back-to-back champions.
15. **Champion Age Trends:**
    - Identify the age ranges where drivers consistently win championships across different decades.
16. **Bonus Challenge (Optional)**
    - Predict the future team of a driver based on past team transitions and transfer trends.

Deliverables

1. **Predictive Model:** A machine learning model capable of forecasting outcomes.
2. **Data Storytelling Report:** Analysis, feature engineering, and insights supported by visualizations.
3. **Presentation:** A clear and engaging summary of findings and predictions.(Be sure to include all relevant details when providing your answer.)

Conclusion

The **Data Science Premier League Hackathon** challenges participants to push the boundaries of data-driven prediction, blending analytics, machine learning, and visualization to make compelling forecasts. the competition demands precision, strategy, and creativity.

Are you ready to step into the analyst's hot seat and make game-changing predictions? Let the **Data Science Premier League** begin!