Capstone Proposal 1:

**[Predicting Formula 1 Race Outcomes with FastF1 Data]**

**Business Understanding:**

**Problem Statement:** This project aims to predict Formula 1 race outcomes, specifically the pole position, podium finishes, and fastest laps, using historical timing and telemetry data.

**Relevance:** Formula 1 is not only a sport but also a high-stakes business where split-second decisions can have substantial financial implications. Better predictions can enhance team strategies and fan engagement.

**Industry Domain:** This applies to the sports analytics domain, with a focus on motor racing.

**Target Audience:** The primary audience will be Formula 1 teams, strategists, and analysts, as well as the broader community of F1 enthusiasts and sports analysts.

**Impact:** By accurately predicting race outcomes, teams can optimize race strategies, and media outlets can improve their coverage and analysis.

**Research:** Exploration of existing work on race prediction models, performance analysis of drivers and cars, and other statistical analyses in motorsports will be foundational to this project.

**Data Understanding:**

**Data Collection:** <https://www.kaggle.com/datasets/juampimgi/formula-1-season-2023> provides access to a wealth of Formula 1 timing and telemetry data, including lap times, tire usage, and weather conditions.

**Context:** The raw data comes from the official Formula 1 data streams, ensuring accuracy and relevance.

**Comparative Analysis:** Other predictive models and analytical methods used in motorsports will be explored to identify gaps and opportunities for innovation.

**Data Preparation:**

**Data Storage:** Data will be extracted from <https://www.kaggle.com/datasets/juampimgi/formula-1-season-2023> and stored in CSV format

**Data Types:** We anticipate a mix of continuous (lap times, tire wear) and categorical data (driver, team, circuit).

**Preprocessing Steps:** Anticipated steps include cleaning missing data, normalizing lap times across different circuits, and encoding categorical variables.

**Challenges:** The intricacies of telemetry data and its high dimensionality will pose significant preprocessing challenges.

**Data Volume:** Aiming for a dataset encompassing multiple seasons for robustness—approximately several tens of thousands of laps.

**Visualization:** Visualizing driver performance, tire degradation patterns, and the impact of weather conditions on race outcomes.

**Modeling:**

**Techniques:** Time series analysis for lap time predictions, classification algorithms for podium finishes, and clustering for driver performance patterns.

**Target Variable:** The target variables will be the qualifying position, race position, and fastest lap times.

**Baseline Model:** A linear regression model for time predictions and logistic regression for classification tasks.

**Problem Type:** Both regression (time predictions) and classification (race outcomes).

**Evaluation:**

**Success Metrics:** Accuracy for classification tasks, RMSE (Root Mean Square Error) for time predictions, and the precision of clustering outcomes.

**MVP:** A model that accurately predicts the top 10 qualifying positions based on historical data.

**Stretch Goals:** Expanding the model to predict entire race outcomes and developing a live-prediction model that can update during race weekends.

**Deployment**

**Reporting:** Results will be shared through an interactive dashboard that visualizes predictions vs. actual outcomes.

**Deployment Plan:** The goal is to develop a web application that allows users to input variables and receive race outcome predictions.

**Tools/Methodologies**

**Python Libraries:** FastF1 for data collection, Pandas for data manipulation, NumPy for numerical operations, Matplotlib/Seaborn for visualization, Scikit-learn for modeling.

**Algorithms:** Linear Regression, Logistic Regression, k-Means Clustering, ARIMA for time series prediction.

**Environment:** Analysis will be conducted locally with potential use of cloud-based resources for increased computational power.

**Data Storage:** Initially on local machine, considering cloud storage (AWS S3) for larger datasets or if web deployment is pursued.