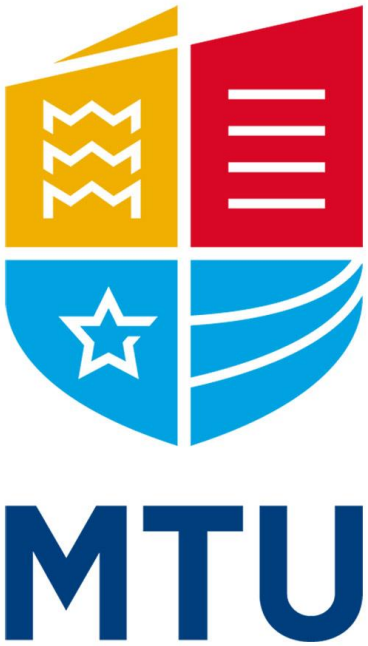

Student Name: PREM VIKAS

Student No: R00241672

For the module Data9005 as part of the

Master of Science in Data Science and Analytics, Department of Mathematics, 2023/24

Declaration of Authorship

I, PREM VIKAS, declare that the work submitted is my own.

- I declare that I have not obtained unfair assistance via use of the internet or a third party in the completion of this examination
- I have not used ChatGPT, an AI chatbot or other similar software in this assignment
- I have acknowledged all main sources of help
- I acknowledge that the Academic Department reserves the right to request me to present for oral examination as part of the assessment regime for this module.
- I confirm that I have read and understood the policy and procedures concerning academic honesty, plagiarism and infringements
- I understand that where breaches of this declaration are detected, these will be reviewed under MTU (Cork) policy and procedures concerning academic honesty, plagiarism and infringements, as well as any other University regulations and policies which may apply to the case. I also understand that any breach of academic honesty is a serious issue and may incur penalties.
- EXAMINATION/ASSESSMENT MATERIAL MAY, AT THE DISCRETION OF THE INTERNAL EXAMINER, BE SUBMITTED TO THE UNIVERSITY'S PLAGIARISM DETECTION SOLUTION
- Where I have consulted the published work of others, this is always clearly attributed
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this work is entirely my own work

Signed: PREM VIKAS

Date: 11-03-2024

Table of Contents

Table of Figures

1. Introduction

This project explores the intricate and data-intensive world of Formula 1 racing, a domain where high-speed competition meets advanced engineering. It aims to reveal hidden patterns and insights within the sport by employing advanced data visualization techniques. The focus is on dissecting driver performances, race outcomes, and the unique characteristics of each track to deepen the understanding for fans and analysts alike.

By analysing factors such as driver consistency, seasonal improvements, and the impact of track configurations on race dynamics, the project seeks to provide a fresh perspective on the elements contributing to success in Formula 1. The visualizations crafted aim to highlight the complex relationship between individual talent and technical strategy, offering an enriched view of the sport's multifaceted nature.

Ultimately, this endeavour is designed to enhance appreciation for Formula 1's detailed landscape, bridging data-driven insights with the sport's rich, competitive essence, thereby offering a nuanced exploration tailored to both enthusiasts and professionals.

1.1 Source of Concept

This report originated from my deep interest in Formula 1, a sport celebrated for its speed, innovation, and strategy. Beyond the excitement, I was drawn to the underlying data, seeking to unearth hidden narratives such as driver tactics, track impacts, and team decisions. This report reflects my journey, merging personal curiosity with insights from F1's online discourse to craft visualizations that illuminate the sport's complexities. Without replicating existing analyses, I've blended public commentary with personal investigation, creating a unique exploration that delves into the multifaceted world of Formula 1 through a fresh analytical lens.

1.2 Data Overview.

In this project, we utilized a comprehensive Formula 1 dataset from Kaggle, originally from the Ergast Developer API. This dataset, split into multiple CSV files, provides a deep dive into various aspects of the sport, covering race tracks, driver profiles, team details, race outcomes, lap times, pit stops, and championship standings across seasons. Each file offers insights into different segments of Formula 1, from individual performances to overall season dynamics.

Prior to analysis, the data was rigorously pre-processed, including anomaly cleaning and dataset merging, to ensure accuracy and facilitate detailed visualizations. This foundational work was critical in crafting an extensive exploration of the sport's intricate dynamics, enabling informed and nuanced analysis through well-structured visualizations.

1.3 Literature Review.

In reviewing the Formula 1 dataset visualizations on Kaggle, most are limited to basic bar and pie charts, providing straightforward insights such as driver standings and race outcomes. In contrast, our project expands significantly beyond this foundational approach, employing a variety of advanced visualization techniques, including GIS mapping, bubble graphs, line charts, and 3D graphs. This diverse toolkit allows for a deeper, multidimensional exploration of the dataset, revealing the global

distribution of circuits, the dynamic performance of drivers over seasons, and intricate race strategies. By transcending conventional visualization methods, our project offers a more detailed and nuanced narrative of Formula 1, illustrating complex patterns and trends that provide a richer understanding of the sport's intricacies compared to the simpler analyses typically found on Kaggle.

2. Visualization.

The visualizations in this project were developed using Python, leveraging libraries such as matplotlib.pyplot for foundational graphs, and plotly.graph_objects and plotly.express for advanced, interactive charts. These tools facilitated the creation of diverse visual formats, from basic bar charts to intricate 3D maps, enhancing the storytelling aspect of Formula 1 data.

We also employed datetime for precise time-series analysis, crucial for interpreting race timings and historical trends. Interactive elements were introduced through ipywidgets and interact, allowing users to engage directly with the data, exploring different scenarios and outcomes. The IPython.display module seamlessly integrated these dynamic components into the Jupyter notebook environment, providing a rich, user-friendly experience.

This approach not only elevates the analysis beyond simple static visuals but also offers a deeper, more engaging exploration of the intricate world of Formula 1 racing.

2.1 Line Chart of Driver Rankings Throughout Selected Formula 1 Season.

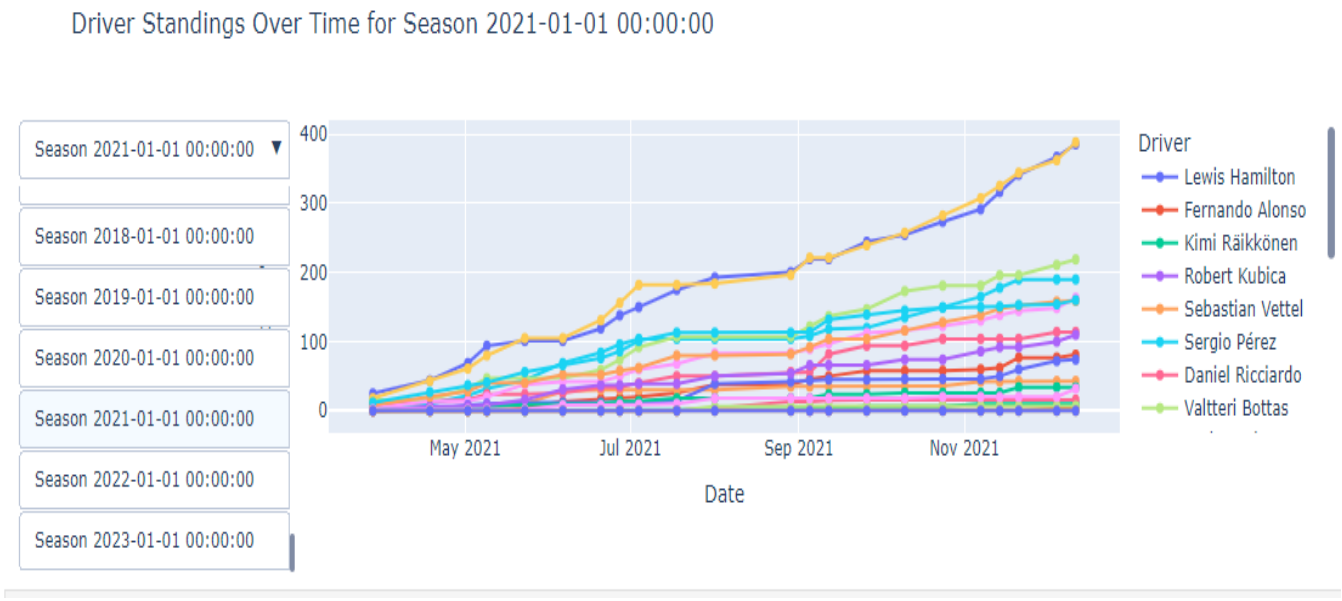

Fig 1. Line chart of drivers rankings

The visualization provides insights into drivers' performance trends, their progress over a season, and comparisons between drivers. This interactive and temporal approach offers a detailed view of the competitive landscape in Formula 1 over time, making complex data more accessible and understandable.

The script uses Plotly to construct an interactive line chart, where each line represents a driver's cumulative points throughout a season. The chart includes an update button for each season, allowing users to switch between seasons and view different datasets. Only the data for the initially selected season is visible by default, enhancing clarity and focus.

The code merges Formula 1 racing data to visualize the cumulative points of drivers across different seasons, allowing for dynamic season-by-season analysis. Initially, it merges driver results with season and personal information, creating a comprehensive dataset (driver_results_full). It then calculates cumulative points for each driver within each season, sorting this data by date for chronological plotting.

2.2 Mapping Driver Victories by Circuit with Scattergeo Visualization.

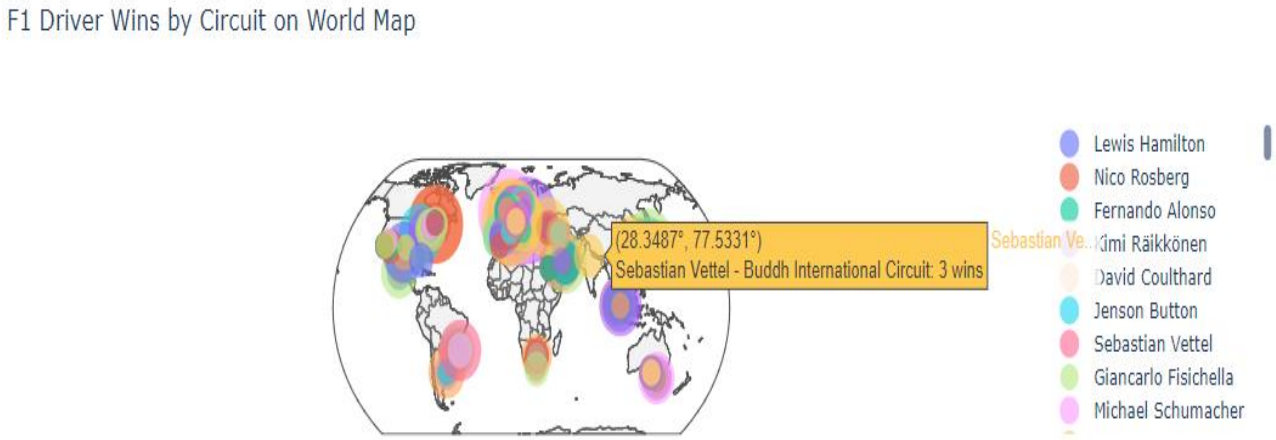

Fig 2. Mapping Driver Victories on geo.

In the visualization phase, a map with markers for each circuit where a driver has won is created. Markers vary in size based on win counts and display detailed information on hover. The result is an interactive world map in Plotly, illustrating each driver's success at various circuits, with legend entries representing individual drivers for easy identification and comparison.

This Python code merges Formula 1 datasets to analyse and visualize drivers' wins per circuit globally. Initially, it integrates race results with drivers and circuits, focusing on victories by filtering for first-place finishes. The wins are then aggregated by driver and circuit, providing a count of how many times each driver has won at each track.

Each circuit's geographical coordinates determine marker positions, while the size reflects the number of wins, enhancing visual distinction. Hover text combines driver names with track names and win counts, offering detailed insights upon interaction.

2.3 Scatter Mapbox Visualization of Quickest Pit Stops at Each Circuit

This visualizes the fastest Formula 1 pit stops for each circuit on a global map. It begins by merging the 'pitstops' dataset with 'races' to add race year and circuit ID to each pit stop entry, then further merges with 'circuits' to incorporate circuit location data (latitude and longitude).

The dataset is then grouped by circuit ID and name, using aggregation to find the minimum (fastest) pit stop time, in milliseconds, for each circuit. This aggregated data is merged back with the 'circuits' dataset to ensure each circuit's name and location are accurately represented.

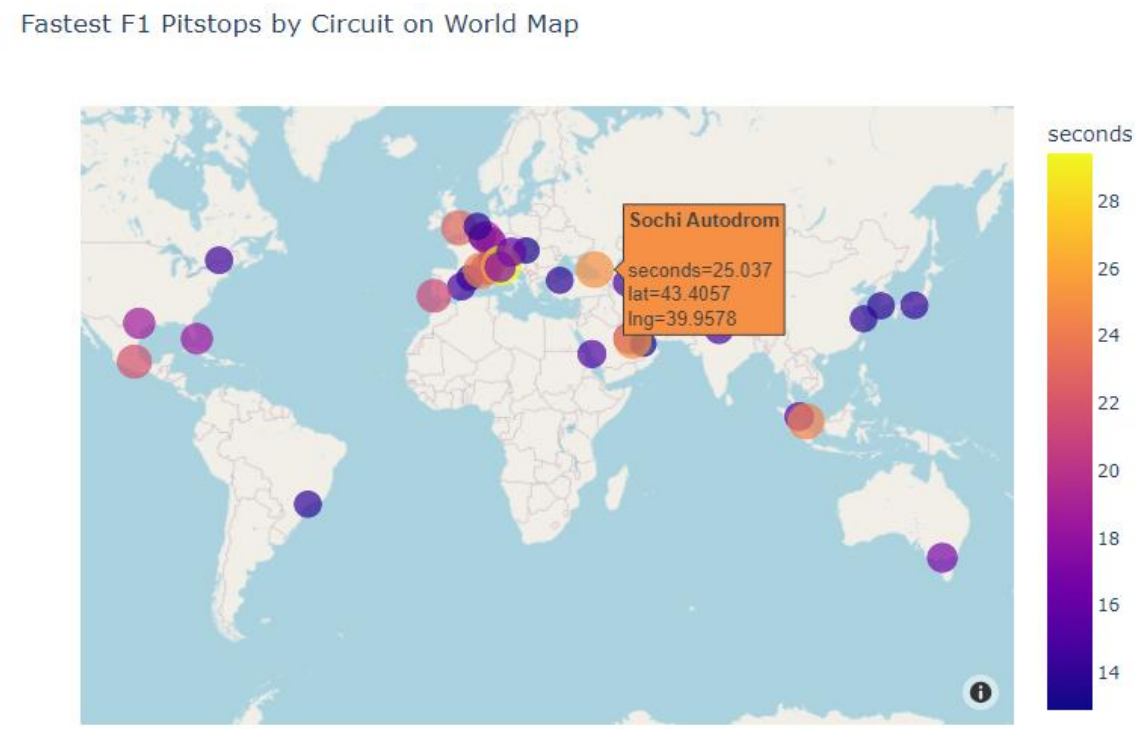

Fig 3. Fastest pitstops for each track

Conversion of pit stop times from milliseconds to seconds enhances readability. Using Plotly Express, the code plots each circuit on a world map as a scatter plot point, where the colour and size of each point are determined by the fastest pit stop time recorded at that location. Points are coloured based on how quick the pit stops were, allowing easy identification of circuits with exceptionally fast or slow stops.

The interactive map, styled with 'open-street-map', enables users to hover over each point to see the circuit name and its record for the fastest pit stop, effectively combining geographical and performance data for insightful visualization. The scatter mapbox is very detailed orientated one can also view the track layout when zoomed in like below.

Fig 4. Interactive map with track layout

This approach aimed to reveal geographical patterns in pit stop efficiency and how track location influences team strategies and performance. The GIS mapping allowed us to correlate pit stop data with specific circuits, providing a visual understanding of how location-based factors might impact the dynamics of race strategies and outcomes, spatial analysis by investigating the **distribution comparision** of pit stop durations plotted on a GIS map for each Formula 1 track.

2.4 Bar Chart Representing a Driver's Total Career Achievements.

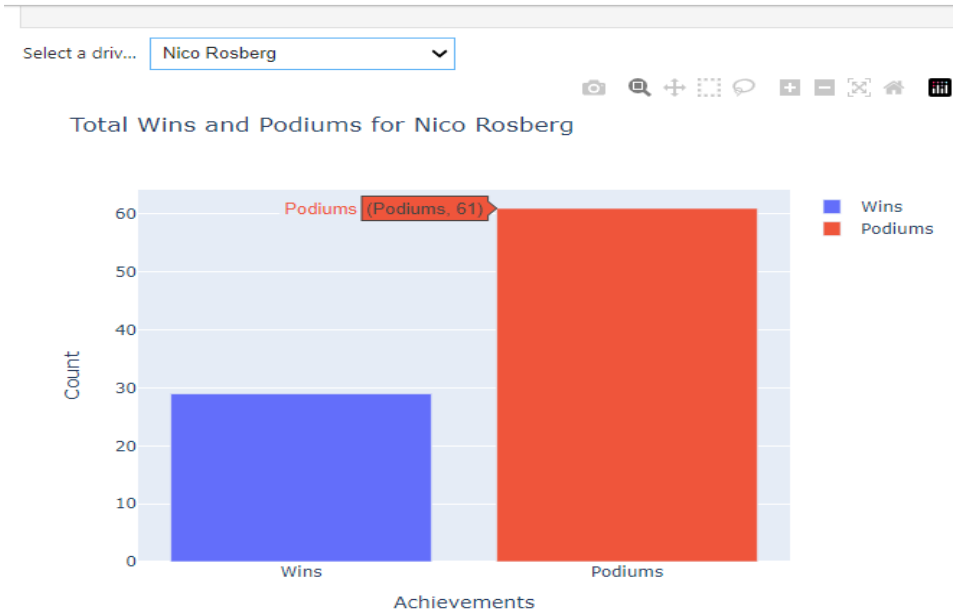

Fig 5. Bar chart of driver achievements.

A bar chart is then generated using Plotly: one bar represents the total wins, and the other represents total podiums. The chart dynamically updates based on the selected driver, displaying their name in the title and adjusting the bars accordingly.

This Python code integrates Plotly and ipywidgets to create an interactive dashboard displaying Formula 1 drivers' total wins and podium finishes. The drivers DataFrame is enhanced with a full_name column by concatenating drivers' forenames and surnames. A dropdown menu is then constructed from these full names, allowing users to select a driver, which also highlight the **novel** of this report making the visualization more user friendly and interactive for user. The methods and ideologies used to plot visualizations in this report is also sophisticated and dynamically updating plots

Upon selection, the update_plot function triggers. It identifies the chosen driver by referencing their unique driverRef, then filters the driver_standings DataFrame to gather only the selected driver's data. The function calculates total wins (first-place finishes) and total podiums (top three finishes).

2.5 Geographic Visualization of Seasonal Race Locations.

The below visualization is implemented by an interactive Geographic Information System (GIS) plot visualizing Formula 1 race circuits for a selected season. Utilizing Plotly's Scattermapbox, the map displays race locations (circuits) marked by their geographical coordinates (latitude and longitude). Users can select a specific Formula 1 season using a dropdown menu, dynamically updating the map to show only races from that season.

The **complexity** and novelty in this implementation lie in the integration of interactive widgets with GIS plotting capabilities. The update_gis function, triggered by changes in the dropdown menu, filters the races_ordered DataFrame to obtain data for the selected season, including circuit locations and round numbers. The hover text combines round information and circuit names, providing context-sensitive information as users interact with the map points.

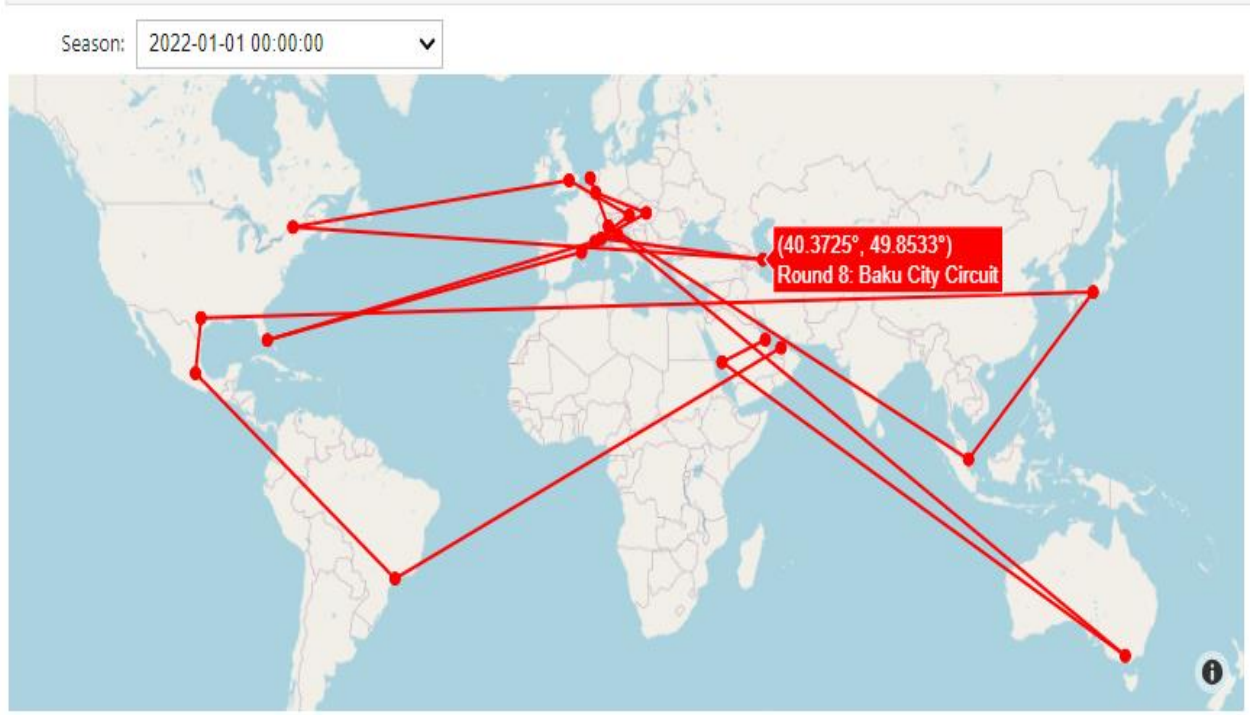

Fig 6. Geographic Visualization for race rounds.

A particularly complex aspect of this implementation is the dynamic calculation and application of map centering based on the average latitude and longitude of the season's circuits, ensuring the map's focus adapts to display all relevant races optimally. The use of go.Scattermapbox within an interactive Python environment illustrates a sophisticated application of data visualization, enhancing user engagement and providing insightful geographical perspectives on Formula 1 seasons.

2.6 3D Bubble Chart Depicting Racing Incidents per Track.

The below 3D bubble chart offers a comprehensive view, allowing for the analysis of spatial patterns and trends in racing incidents across global F1 tracks, thus making complex datasets understandable and visually engaging.

A 3D bubble chart is used to visualise the number of racing incidents at different Formula 1 circuits. Using Plotly's go.Scatter3d, it maps each circuit's geographical location to a point in three-dimensional space, where longitude and latitude form the X and Y axes, and the Z axis represents the total number of accidents reported at that location.

The size of each bubble marker corresponds to the number of accidents, providing an immediate visual indication of the frequency of incidents at each circuit. This size mapping is achieved through the size attribute of the markers, adjusted relative to the maximum number of accidents for clarity and scaled using the sizemode and sizeref parameters.

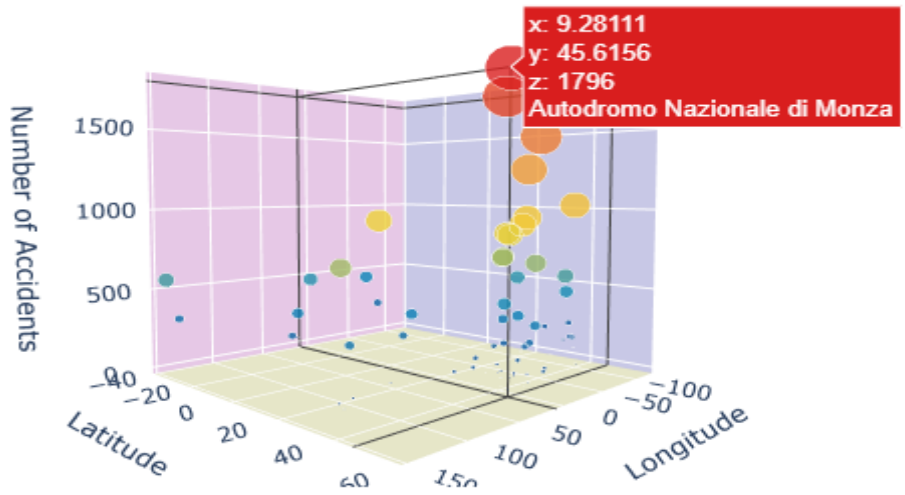

Fig 7. 3D Bubble graph of racing incidents.

Color is also used to differentiate between circuits based on the number of incidents, with a gradient applied through the colorscale attribute to enhance visual differentiation. The chart includes hover text displaying the name of each circuit and the respective data points, improving interactivity and information clarity.

3. Design Considerations/Choice.

Geographical and Advanced Visualization Techniques: Prioritizing GIS allowed me to map Formula 1 data geographically, highlighting circuit locations and their historical significance in the sport. This choice was grounded in the understanding that spatial context significantly enriches the narrative around race outcomes and strategies. Additionally, employing complex visualizations like 3D bubble charts, line charts, scatter plots, and bar charts aligned with the dataset's multifaceted nature. These formats were chosen based on Nathan Yau's principles, which advocate for selecting visualizations that most effectively convey the underlying data patterns and stories, enhancing clarity and insight.

User Interaction: Implementing dropdown menus and interactive graphs was a deliberate choice to improve user engagement and comprehension. By allowing users to select specific datasets, like seasons or drivers, the interface provides a more tailored exploration experience. This approach facilitates a deeper understanding as users can dynamically interact with the data, uncovering details that static charts could not easily reveal.

Data Visualization and Color Scheme: In visualizing key concepts such as the number of rounds per season and driver achievements, I aimed to make the information accessible and meaningful. The deliberate choice of colors to reflect real F1 team colors in bar charts is an example of this, enhancing the visual appeal and relatability of the graphs. This decision aligns with the best practices advocated by experts in data visualization, who stress the importance of color in encoding information and creating an intuitive user experience. By mirroring actual team colors, fans can quickly associate data points with the corresponding teams, deepening the connection between the viewer and the information presented.

In our analysis, the scatter plot for winning teams over time illustrates the changing dominance within Formula 1, identifying trends of superiority or **pattern trends** in team performance across seasons. This visualization helps in understanding the impact of technological advancements and strategic changes on team success.

The bar chart displaying constructor points per race reveals the consistency and competitiveness of teams, highlighting those with steady performances versus those more erratic. This aids in evaluating team strategies and their success across various races.

Although these visualizations are not detailed in the report, they are crucial in the code section for analyzing Formula 1 trends and team dynamics, offering insights into historical patterns and current competitiveness in the sport.

4. Future work.

Given more time, my work could be significantly expanded to delve deeper into Formula 1 analytics, providing a richer and more nuanced understanding of the sport. Firstly, I would develop a dedicated database encompassing detailed track metrics for each season. This would allow for the visualization of specific track characteristics and how they correlate with individual driver performances, highlighting which drivers excel on particular types of circuits, such as high-speed tracks versus those requiring high downforce.Furthermore, I would focus on visualizing historical rivalries, tracking their evolution over seasons. This would involve creating dynamic visual comparisons to showcase head-

to-head performances, changes in team dynamics, and shifts in competitive edge, thus bringing to life the dramatic narrative of F1 rivalries.

To enhance the analytical depth and presentation quality, I would incorporate principles from renowned data visualization experts such as Edward Tufte, known for his minimalist approach focusing on clarity and information density, and Alberto Cairo, who emphasizes the balance between functionality and aesthetics in data storytelling. Incorporating their methodologies would improve the clarity, engagement, and interpretability of the visualizations.

Word Count= 2390.

References.

1. https://www.kaggle.com/code/sanketdevhare98/formula-1-data-visualization by **SANKET DEVHARE.**
2. https://www.kaggle.com/code/umairziact/formula-1-visuals-eda by **UMAIR ZIA.**
3. https://www.youtube.com/watch?v=UO98lJQ3QGI&list=PL-osiE80TeTvipOqomVEeZ1HRrcEvtZB **by Corey Schafer.**
4. https://www.youtube.com/watch?v=4O_o53ag3ag&t=2s **by Rob Mulla.**
5. https://r.geocompx.org/spatial-class.html#vector-data**.**
6. https://ebookcentral.proquest.com/lib/munster/reader.action?docID=1158630&query=data+points**.**