Analyzing Characteristics of Formula One Circuits Using Spatiotemporal Data

Taylor Fourier, B.S., San Diego State University
Chloe McCormick, B.S., San Diego State University
Madison Nafarrete, B.S., San Diego State University
Rachel Schavel, B.S., San Diego State University

I. Abstract

Track attributes play a pivotal role in Formula One, exerting influence on race outcomes. Leveraging racer and event data obtained from an open-source repository on GitHub, spanning race results from 2013 to the present, alongside shapefile data detailing the geographical layout and features of each circuit, this study employs correlation analyses, and heatmaps to unveil potential relationships between track geography and various race aspects. Through diverse analysis, including examinations of overtaking patterns, geography-based strategic approaches, and racing outcomes, our research aims to contribute to a comprehensive understanding of how track configurations and attributes interplay with decision-making within the Formula One ecosystem.

Keywords: Formula One, event data, correlation models, overtakes, spatiotemporal, heatmap

II. Introduction

Formula One, commonly denoted as F1, represents the foremost motor racing competition on a global scale. Since its inception in 1950, it has garnered widespread praise, showcasing a diverse array of 77 unique Grand Prix circuits spanning 34 countries[1]. In contrast to invasion sports, which revolve around structured gameplay and team-based possession dynamics[2], Formula One distinguishes itself through its emphasis on the intricate interplay between car engineering and the drivers' mastery of vehicle control. Furthermore, the varied designs of Formula One circuits introduce external factors that shape the dynamics of athlete and vehicle interaction, presenting unique challenges for each race.

Academically accredited spatiotemporal research mainly focuses on invasion sports looking at player movements relative to various plays[2][3]. However, with Formula One, we endeavor to adopt a more geographical perspective using this analytical approach, examining the influence of circuit attributes on race dynamics in the year 2018.

The paper is structured as follows: Section 3 delineates the datasets utilized in this analysis. Section 4 provides an in-depth exposition of our methodology for data collection, organization, and processing. In Section 5, we conduct an analysis of visualizations derived from our methodology. Section 6 serves as the concluding section, acknowledging limitations and other research areas that could further this discussion. Finally, Section 7 acknowledges and highlights the contributors to this paper.

III. About the Datasets

Formula1_2018Season_RaceResults.csv

The dataset Formula1_2018Season_RaceResults.csv originates from a publicly accessible GitHub repository, specifically attributed to the user *toUpperCase78*, which sourced its data from the official Formula One website. The precise methodology employed for data collection and sorting within this repository remains unknown. Nonetheless, this dataset encompasses a range of pertinent attributes, notably including the geographical location of the race track, positional rankings, unique identifiers for the participating cars, driver names, affiliated teams, starting grid positions, laps completed, race times, points accrued, and respective fastest lap records attained during the races.

Comprising 421 observations, this dataset is arranged in chronological order according to the sequence of the twenty-one Formula One Grand Prix occurring throughout the 2018 season starting with the opening race held in Australia and culminating with the concluding event in Abu Dhabi.

F1-circuits.geojson

The dataset F1-circuits.geojson originates from a publicly accessible GitHub repository, specifically attributed to the user *backinger*, which sourced its data from the official Formula One website. The precise methodology employed for data collection and sorting within this repository remains unknown. This dataset has 42 different tracks that include columns: ID, Location, Name, Debut Year, Length, Altitude, and Coordinates. In our analysis, we only used the 21 tracks that were raced on during the

2018 racing year. This included eleven tracks in Europe, five tracks in Asia, three tracks in North America, one track in South America, and one track in Oceania.

2018Overtakes-Sheet1.csv

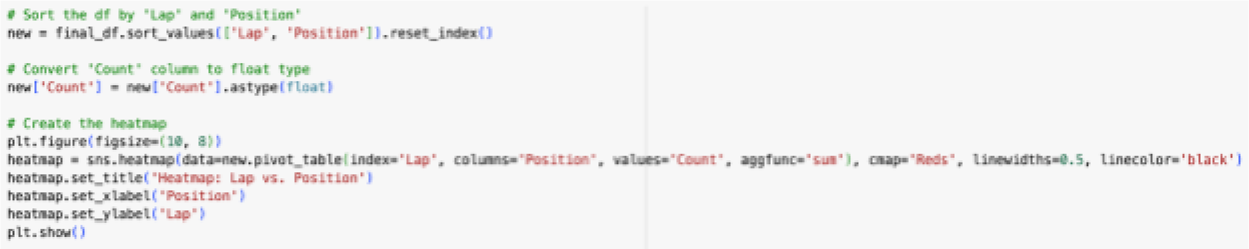
The Formula One Reddit thread, r/formula1, features a post by *catchingisonething*, presenting overtake data from each season since 1994. Specifically focusing on the 2018 season, the dataset comprises 750 recorded overtakes and is structured in tabular format, detailing the race session, lap number of the overtake, car position during the overtake, the athlete performing the overtake, the athlete being overtaken, indication of whether the overtake was recorded live, and the turn where the overtake occurred. Similar to the previous data files, this dataset follows a sequential order by Grand Prix throughout the season.

Due to the limited availability of overtake data online, the author employed several methods to gather and organize the dataset. Initially, overtakes were counted using a lap chart and pit stop data. Subsequently, the races were rewatched to add any missed overtakes, with a focus on passing maneuvers during laps and instances of back-and-forth passing on the same lap. Overtakes occurring on the first lap or after restarts were excluded due to their complexity. Only overtakes made for position were considered, with those occurring during pit stops, spins, or other incidents being omitted. Additionally, the presence of overtakes on TV and their specific location on the track were recorded but were not used in this analysis. While acknowledging the possibility of errors, the overall accuracy of the data regarding on-track passing is regarded as reliable for purposes of this project .

IV. Methodology for Analysis

Python was used for all of the coding, and a variety of libraries were used to read, interpret, and visualize our three datasets. In particular, Pandas was used to read the raster data on overtakes and the racer data; GeoPandas was used to read the shapefile. Visualization was a key factor in answering the main research question, so a majority of the plots and graphs were generated with the Matplotlib library.

Creating a Heatmap



```python
# Sort the df by 'Lap' and 'Position'
new = final_df.sort_values(['Lap', 'Position']).reset_index()

# Convert 'Count' column to float type
new['Count'] = new['Count'].astype(float)

# Create the heatmap
plt.figure(figsize=(10, 8))
heatmap = sns.heatmap(data=new.pivot_table(index='Lap', columns='Position', values='Count', aggfunc='sum'), cmap='Reds', linewidths=0.5, linecolor='black')
heatmap.set_title('Heatmap: Lap vs. Position')
heatmap.set_xlabel('Position')
heatmap.set_ylabel('Lap')
plt.show()
```

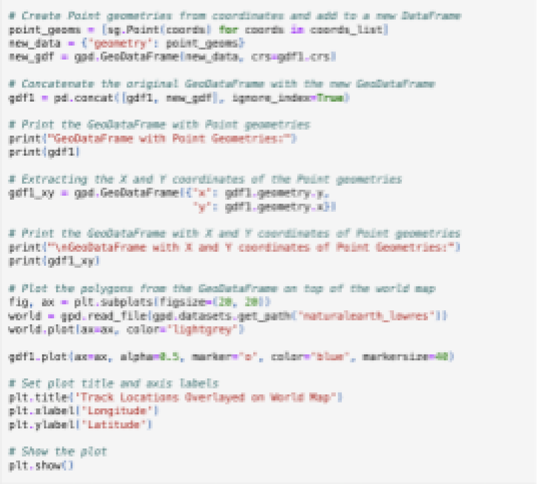
Figure 1: Code for creating heatmap

The integration of overtake data played a pivotal role in constructing heatmaps aimed at revealing how track configurations influence racing outcomes. This dataset provided valuable insights into the specific turns where overtakes occurred, which were then plotted onto maps extracted from shapefiles using Pandas and Matplotlib. By sorting the 'Overtakes' dataset based on lap and position columns, a heatmap was created. The index of the data frame was reset so it became sequential starting from 0. Following this, the count column was converted into a float type to ensure the numerical operations could be performed on the column correctly.

Mapping the Tracks

The initial approach to addressing the research question centered on examining the geographical positioning of the circuits raced in the 2018 season. Shapefiles served as a fundamental tool for this task, providing essential spatial information about track locations. However, challenges arose during the reading of these files, particularly on macOS systems, with occasional success on Windows platforms. Despite this, progress was achieved by extracting coordinates from the GeoDataFrame, facilitating the generation of points for integration onto a large-scale map.

The constructed map utilized GeoPandas with a Natural Earth projection. Matplotlib further enhanced the map by customizing labels and points' appearance.



```python
# Create Point geometries from coordinates and add to a new DataFrame
point_geoms = [sg.Point(coords) for coords in coords_list]
new_data = {'geometry': point_geoms}
new_gdf = gpd.GeoDataFrame(new_data, crs=gdf1.crs)

# Concatenate the original GeoDataFrame with the new GeoDataFrame
gdf1 = pd.concat([gdf1, new_gdf], ignore_index=True)

# Print the GeoDataFrame with Point geometries
print("GeoDataFrame with Point Geometries:")
print(gdf1)

# Extracting the X and Y coordinates of the Point geometries
gdf1_xy = gpd.GeoDataFrame({'x': gdf1.geometry.y,
                            'y': gdf1.geometry.x})

# Print the GeoDataFrame with X and Y coordinates of Point geometries
print("\nGeoDataFrame with X and Y coordinates of Point Geometries:")
print(gdf1_xy)

# Plot the polygons from the GeoDataFrame on top of the world map
fig, ax = plt.subplots(figsize=(20, 20))
world = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))
world.plot(ax=ax, color='lightgrey')

gdf1.plot(ax=ax, alpha=0.5, marker='o', color='blue', markersize=40)

# Set plot title and axis labels
plt.title('Track Locations Overlayed on World Map')
plt.xlabel('Longitude')
plt.ylabel('Latitude')

# Show the plot
plt.show()
```

Figure 2: Code for circuit locations projected on world map

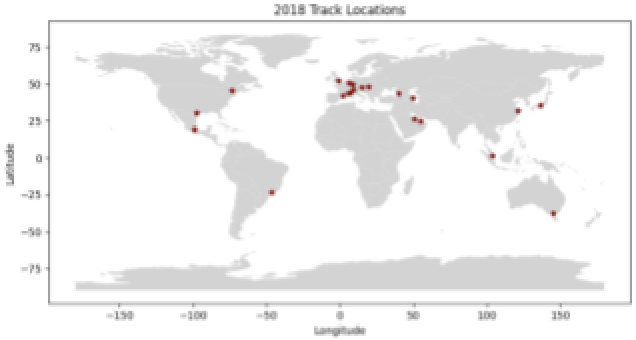
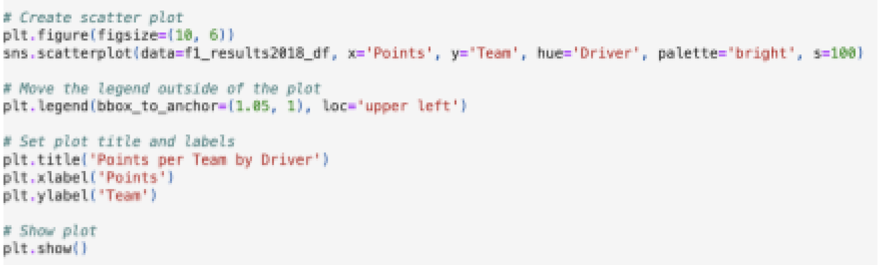


Figure 3: 2018 Track Locations on world map output

Alongside the world map, individual circuit maps of the 21 tracks highlighted during the season were generated also with GeoPandas. These maps provided crucial context regarding the spatial distribution of the circuits for the related race results data and overtake data, enabling subsequent analyses.

Summary Statistics for Racer Data

Before proceeding with summary statistics and integrating racer data with other datasets, various filtering steps were essential. Groups were organized based on track location to compare differences in fastest lap times across tracks. Additionally, filtering was applied to teams and individual drivers to assess their impact on point-related outcomes. Formula One's points system allocates points based on race finishing positions, including an extra point for the fastest lap within the top 10 finishers. It's noteworthy that data on racers who did not finish or were disqualified were excluded from lap time and position analyses. A scatterplot was created to compare points per team and per driver for the 2018 season, highlighting top-performing teams such as Mercedes, Ferrari, and Red Bull Racing. Visualization of racer data was facilitated through the Pandas, Matplotlib, and Seaborn libraries, while statistical analysis relied on the SciPy library.



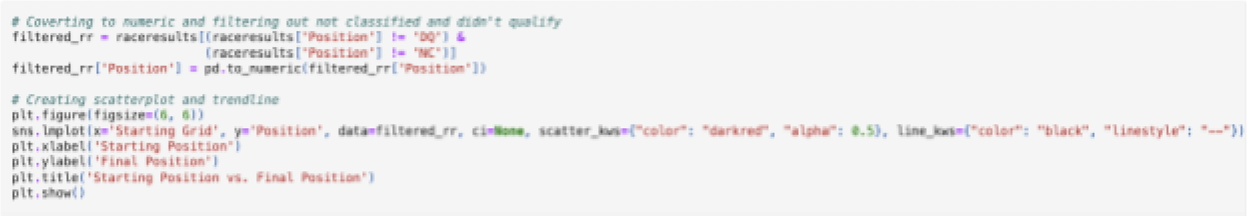
```
# Create scatter plot
plt.figure(figsize=(10, 6))
sns.scatterplot(data=f1_results2018_df, x='Points', y='Team', hue='Driver', palette='bright', s=100)

# Move the legend outside of the plot
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')

# Set plot title and labels
plt.title('Points per Team by Driver')
plt.xlabel('Points')
plt.ylabel('Team')

# Show plot
plt.show()
```

Figure 4: Code creating a plot for points allocated to each team by driver



```
# Coverting to numeric and filtering out not classified and didn't qualify
filtered_rr = raceresults[(raceresults['Position'] != 'DQ') &
                          (raceresults['Position'] != 'NC')]
filtered_rr['Position'] = pd.to_numeric(filtered_rr['Position'])

# Creating scatterplot and trendline
plt.figure(figsize=(6, 6))
sns.lmplot(x='Starting Grid', y='Position', data=filtered_rr, ci=None, scatter_kws={"color": "darkred", "alpha": 0.5}, line_kws={"color": "black", "linestyle": "--"})
plt.xlabel('Starting Position')
plt.ylabel('Final Position')
plt.title('Starting Position vs. Final Position')
plt.show()
```

Figure 5: Code for creating position scatterplot

To delve deeper into the relationship between lap times and points, data was filtered by track location, unveiling disparities in top lap times across tracks.

A scatterplot with a trendline was utilized to explore the correlation between starting positions and final positions, providing valuable insights into race dynamics. Furthermore, a one-sample proportion test conducted using SciPy aimed to discern whether a specific team consistently achieved significantly faster lap times throughout the season compared to others.

V. Results of Analysis

To delve deeper into the insights gained from the methodology, the intricate relationship between track configurations and race dynamics becomes more apparent through the upcoming visualizations. Building upon the initial groundwork done in the methodology, the analysis explores the nuanced interactions between circuit attributes and various facets of race performance. Through examination and visualization, this section aims to investigate the underlying patterns and correlations that shape Formula One racing outcomes.

In Formula 1 racing, grid positioning plays a crucial role in shaping a driver's performance during a Grand Prix race. These positions are typically determined through multiple qualifying sessions, where drivers compete for the best starting spot based on their fastest lap times. This positioning can hold significant influence over a driver's chances for success during the race, especially considering the anatomy of a circuit can change the opportunities for passing.

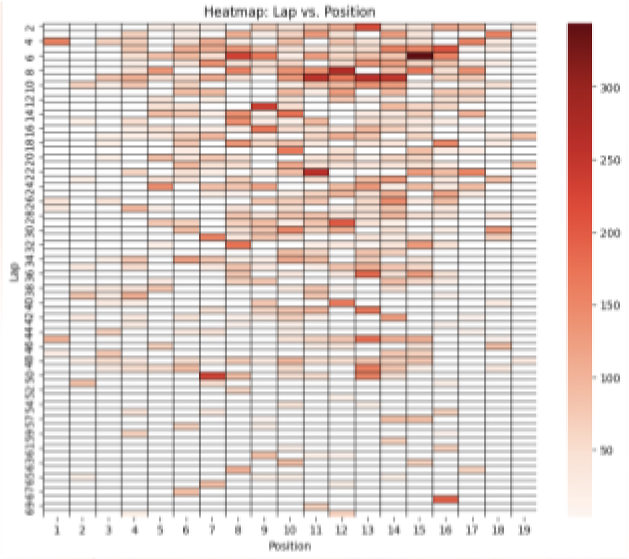


Figure 6: Heatmap Lap vs Position

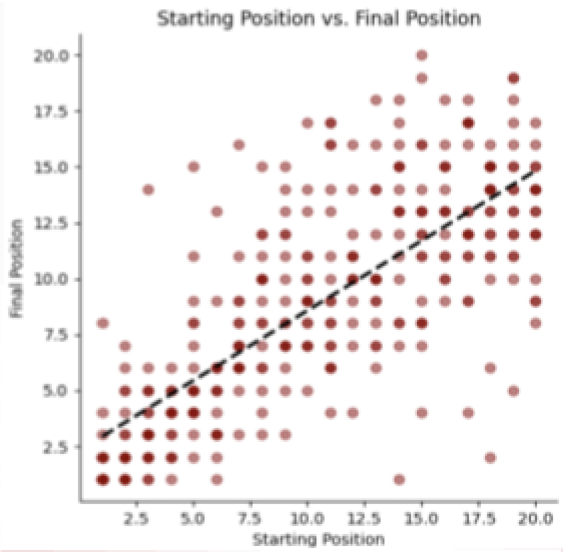


Figure 7: Position Scatterplot

Our heat map analysis reveals a consistent trend: drivers starting from intermediate positions tend to face more overtakes during the initial laps of the race across all tracks. Moreover, as the race progresses, the number of overtakes decreases. In contrast, competitors beginning from higher grid positions experience fewer overtaking moves overall. This correlation parallels the observations in the scatterplot, showing a positive relationship between starting positions and final standings. Specifically, drivers who start from higher grid positions often maintain their advantageous spots throughout the race.

Moreover, a detailed examination of individual track geometries utilizing shapefile data has unveiled compelling insights into the correlation between track layout and lap times. As illustrated in Figure 8, a comparative analysis between the Red Bull Ring in Spielberg, Austria, and the Baku City Circuit in Azerbaijan highlights notable distinctions. Austria's Red Bull Ring stands out for resulting in the season's fastest recorded top lap time, whereas Azerbaijan's Baku City Circuit records

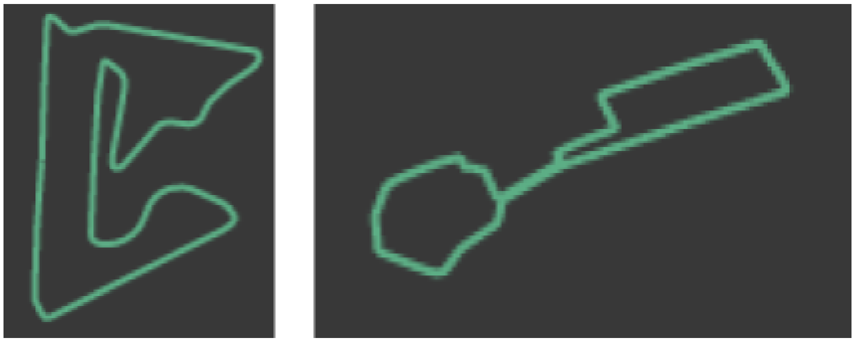
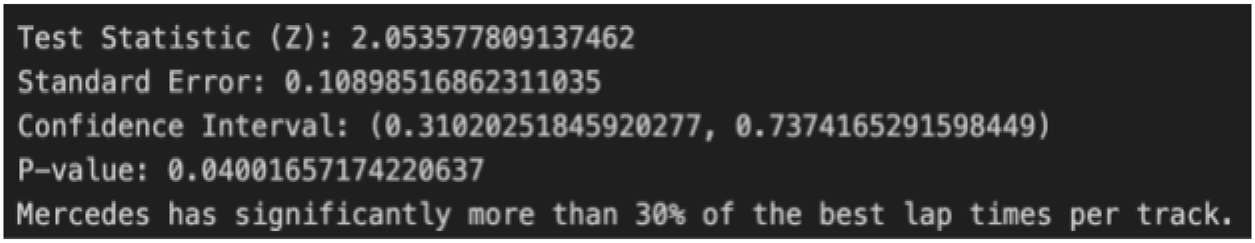


Figure 8: Red Bull Ring (Left), Baku City Circuit (Right)

the slowest. This discrepancy in lap times can be attributed to the inherent characteristics of each track's layout.

The track shapes suggest that circuits with longer straightaways, exemplified by the Red Bull Ring, are conducive to quicker lap times. These lengthy straights enable drivers to maintain higher speeds, thus facilitating overtaking maneuvers. In contrast, circuits characterized by more intricate layouts, such as the Baku City Circuit, present challenges for drivers due to the presence of numerous curves that necessitate slower speeds. Consequently, overtaking becomes more difficult on such circuits.



```
Test Statistic (Z): 2.053577809137462
Standard Error: 0.10898516862311035
Confidence Interval: (0.31020251845920277, 0.7374165291598449)
P-value: 0.04001657174220637
Mercedes has significantly more than 30% of the best lap times per track.
```

Figure 9: Hypothesis Test for Mercedes Lap Times

An exploration of racer statistics, portrayed through scatter plots (refer to Figure 9), clarifies the dominance of prominent teams such as Mercedes, Ferrari, and Red Bull Racing, consistently amassing the highest points totals throughout the Formula 1 season. Leveraging both the filtered fastest lap data by track and the points per team dataset, a one-sample proportion test was conducted utilizing SciPy to ascertain whether Mercedes consistently secured 30% or more of the best lap times per track. Remarkably, the test yielded significant results at the 0.05 alpha level. Referencing an outside independent analysis on reliability for the season[4], Mercedes only had a total of two retirements due to technological issues one being from the Austrian Grand Prix and the second in the Brazilian Grand Prix. In fact, all three of the dominant teams mentioned before had top scores in reliability, thus suggesting that reliability, lessening the time during pit stops, improves the performance of racers.

VI. Concluding Remarks

While this study has shed light on various aspects of Formula One racing dynamics, it is important to acknowledge several limitations and areas for future research. Firstly, the analysis focused exclusively on the 2018 season, and subsequent studies could investigate longitudinal trends across multiple seasons to discern evolving patterns and dynamics. Additionally, the study relied primarily on publicly available datasets sourced mainly from fans of the sport, rather than a comprehensive database from Formula 1 itself or an openly accessible API. Future research could benefit from accessing more extensive and detailed datasets to conduct more thorough analyses.

Moreover, although the study examined overtaking patterns and track geography, it is worth noting that other factors such as car design and engineering, weather conditions, usage of different tire types, and driver behaviors may also influence race outcomes and warrant further investigation. Incorporating a multivariate approach into the analysis could offer deeper insights into how different variables interact.

Another area for analysis involves exploring the decision-making process behind selecting circuits for a given season. While the circuits remained consistent from 2018 to 2019, changes in recent years have occurred, making it intriguing to examine which tracks are more favorable from a business and fan perspective. Track selection seems to be based around social, economic and political contexts in addition to Formula 1 tradition, contracts, and of course, circuit regulation. It would also be interesting to look at the long term effects of what Formula 1 events have on their host cities and/or countries from a tourism standpoint.

In summary, this study enhances our understanding of Formula One racing while emphasizing the need for more accessible statistics to advance research, engagement, and development within the sport. By employing data-driven methods and analysis, future studies can further explore Formula One dynamics and contribute to enhancing the sport's competitiveness, excitement and impact.

VII. About the Authors

Taylor Fourier is a fourth-year student at San Diego State University. She is pursuing her Bachelor of Science in statistics with an interest in business analytics, particularly in the intersection of entertainment and sports. Taylor was one of the main contributors to our paper, providing our abstract and introduction, along with visualizing our data and connecting it to our research question.
Email: taylorfourier@gmail.com

Chloe Mccormick is a fourth-year student at San Diego State University. She is about to finish her Bachelor of Science in Statistics. Chloe’s main interest is data visualizations and analytics, leveraging them to effectively convey information to drive

informed decision-making. Chloe was one of the main contributors to the coding for our research, focusing on filtering the data and combining the ‘overtakes’ dataset with the shapefile.
Email: chloemccorick2021@yahoo.com

Madison Nafarrete is a fourth-year student at San Diego State University. She is pursuing her Bachelor of Science in statistics. Madi was one of the main contributors to the coding for our research, focusing on summary statistics for our racer data and connecting it with our overtake data.
Email: mnafarrete7817@sdsu.edu

Rachel Schavel is a fourth-year student at San Diego State University. She is about to finish her Bachelor of Science in statistics with a minor in marketing. Her main interest is in inferential statistics with consumer behavior to help businesses make predictions and optimize performance. For this research, Rachel contributed by finding some basic summary statistics with our racer data as well as recording the methods used in our research, our analysis, and our limitations in the research.
Email: rachelschavel@gmail.com

VIII: References

Bacinger, T. (2024, January 16). F1-Circuits. GitHub.https://github.com/bacinger/f1-circuits/blob/master/f1-circuits.geojson

[1]Braybrook, R., & Noble, J. (2023, December 28). *Which country has hosted the most F1 races? Tracks with the most grands prix*. Motorsport.com. Retrieved April 23, 2024, from https://us.motorsport.com/f1/news/which-country-hosted-most-f1-races/1056092

[4]Collantine, K. (2018, December 28). *2018 F1 season in statistics: Car performance*. RaceFans. Retrieved April 29, 2024, from https://www.racefans.net/2018/12/28/how-close-was-it-10-charts-revealing-the-teams-performance-in-2018/

Hart, G. (2023, October 30). Racing Pass Motor Sports Analysis. Racingpass. https://racingpass.net/

Hart, G. (2021, May 18). R/formula1 on Reddit: F1 Overtaking database 1994-2020. Reddit. https://www.reddit.com/r/formula1/comments/nf4jkq/f1_overtaking_database_19942020/

[2]Marin, A. (2023, February 15). *A Comprehensive Guide to Spatio-Temporal Analysis in Team Sports*. Medium. Retrieved April 23, 2024, from https://medium.com/@

marin11amf11/a-comprehensive-guide-to-spatio-temporal-analysis-in-team-sports-28f8841b9122

Mccormick, C. (2024, March 19). CHLOEZM/Formula-FUN. GitHub.https://github.com/chloezm/Formula-Fun

Schneider, J. (2018, May 22). Polylines3. Google Drive. https://drive.google.com/file/d/18GSOqRNrb7bSErTRVUOCzwFw-7hJ0E1q/

Terenzio, J. (2022, April 15). Finding the limit: Formula 1 data visualizations and points prediction. Medium. https://medium.com/@julianterenzio/finding-the-limit-formula-1-data-visualizations-and-points-prediction-fb42ecad4729

Yenigun, D. (2022, August 19). Formula1-datasets. GitHub. https://github.com/toUpperCase78/formula1-datasets/blob/master/PreviousSeasons/Formula1_2018Season_RaceResults.csv

⌜3⌝,. (2024, March 5)., - YouTube. Retrieved April 23, 2024, from https://www.nature.com/articles/s41597-019-0247-7