Jameson Cohen
COMPSCI 1710: Visualization

CS1710: Process Book

Final Project: Plan

Team Agreement: I am conducting a joint project with my CS1090a project, in which I am working with three other people. My goal is to construct an interactive web-application-style EDA for that project as my final project for CS 1710. Therefore, I will be working alone in the context of CS 1710 and therefore do not have any other team members with which to form a team agreement. If there are any other details I should provide for this section, please let me know, and I would be happy to do so.

Basic Info:
Project Title: Data Insights and Predictions in Formula One Races (EDA)
Name / Team Name: Jameson Cohen
Email: jamesoncohen@college.harvard.edu

Background and Motivation:
Sports analytics has been a topic of interest to me for quite some time. Last year, I was fortunate enough to take Mark Glickman's sports analytics class (STAT 143), which introduced me to several statistical modeling techniques that allowed me to derive interesting insights and predictions from sports data. At the end of that class, I conducted a final project working with similar Formula One data to the Kaggle dataset discussed below, though that project was focused on analyzing a ranked-choice logit model that had already been implemented by the authors of a paper (see related work). In my CS109a project, my group is hoping to use this data to come up with interesting inferences about the data using various regression models and design a well-tuned prediction model for race outcomes. I believe this is a great project to merge with my CS1710 project because, in conducting an exploratory data analysis (EDA) of the data, I believe Formula One data may be best-understood through an interactive visualization, allowing one to visualize different outcomes (race positions, times, pit stops, etc.) at different race tracks over a range of seasons.

Related Work:
The primary piece of related literature that works with data compiled from the same source (Ergast Developer API) is "Bayesian analysis of Formula One race results: disentangling driver skill and constructor advantage" by Erik-Jan van Kesteren and Tom Bergkamp (2023). While working with similar data, the primary goal of this paper is to apply a ranked-choice logit model to understand whether individual driver skill or team advantages (particularly due to car design) are more important for race outcomes. Our goal in my CS1090a project will be to work with more of the available predictor variables to design models that are suitable for both inference and prediction. And notably missing from the paper cited above is any sort of thorough EDA,

particularly one that allows a user to interact with and visualize the data themselves through a web application!

Data:
The dataset I will be using is the following Kaggle dataset: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020. This data is compiled from the Ergast Developer API and includes several .csv files offering different types of data (constructors, drivers, seasons, results, etc.). When merged together, this data provides us access to several different variables related to Formula One races from 70+ seasons that we can consider for visualizations when creating this EDA web application.

Data Cleanup:
The primary data file we will be working with for both the CS1090a portion of this project and the CS1710 portion of this project is "results.csv." However, as part of the CS1090a project, we will be merging this with a couple of the other files, such as the driver and constructor data, in order to have all of the predictor variables we hope to include in our final dataset. There are several, clear identifying variables that will allow us to conduct this merge cleanly. Since the author of the Kaggle data already seems to have cleaned the data upon compiling it from Ergast, this should be most (if not all) of the cleaning work I should have to undertake. After conducting this cleaning work, I will use the resulting dataset to conduct the web-application-based visualization that will form my CS 1710 project (as well as the EDA portion of our CS 1090a project).

Final Project: Map

Note: As a reminder, I am working on this project alone in conjunction with my group for my CS1090a project on Formula One data. For my project, I plan to create an interactive web-application-based EDA as a part of my group's larger project for CS1090a.

Audience

1) The three options for my audience that I came up with are:
    a) Formula One fans who already follow races and drivers regularly but are not overly familiar with how F1 data can inform team and driver strategy
    b) F1 analysts and professional strategists, such as those who work in the data analytics and strategy departments on each individual F1 team or for the sport at large
    c) Newcomers to F1 with an interest in sports analytics and applying the principles of sports analytics to F1

Chosen target audience: Option b – F1 analysts and professional strategists.

2) The target audience that I chose is option b): F1 analysts and professional strategists. In our CS1090a project, we will likely plan to focus on creating models with our available data that will inform race strategy. Specifically, we are currently planning on modeling out pit stops and analyzing pit stop strategy.

   Therefore, it makes sense that our target audience for this interactive EDA will be people who are already familiar enough with F1 strategy (likely not newcomers or regular fans). These users are deeply interested in extracting actionable insights to refine race strategies, compare team performance, and optimize future decisions based on historical data. They focus on specific variables like driver timing under various conditions and aim to identify patterns that can give their teams a competitive edge. This audience has advanced knowledge of F1, including in-depth familiarity with race dynamics, strategic factors (e.g., tire management, pit stop timing, etc.), and F1 regulations. They should have reasonable (though not overly complex) data visualization literacy, due to how frequently their job requires working with data. Finally, they require a fairly high level of detail in visualizations, often down to milliseconds, position-by-lap progression, and multi-race comparisons. F1 is a sport where tiny differences in strategy and performance (such as on a routine pit stop) can make major differences in outcomes.

3) Focusing on pit stop strategy analysis, these are some of the questions that I believe would be interesting to my audience:
   a) What is the optimal pit stop lap for different circuits based on historical data?
   b) How do pit stop durations differ across circuits?
   c) Which teams demonstrate the highest pit stop consistency, and how does this impact race outcomes?
   d) How does pit stop timing correlate with tire degradation rates and lap times for each circuit?
   e) Are there specific lap windows during which pit stops result in the most significant position gains?
   f) How does pit stop strategy vary by weather conditions, and what are the impacts on race outcomes?
   g) What is the relationship between the number of pit stops and final race positions for different circuits and race conditions?
   h) Which constructors achieve the best balance between pit stop speed and race performance, and how does this evolve throughout a season?
   i) How does pit stop strategy change during different phases of a race, such as the first, middle, and final laps?

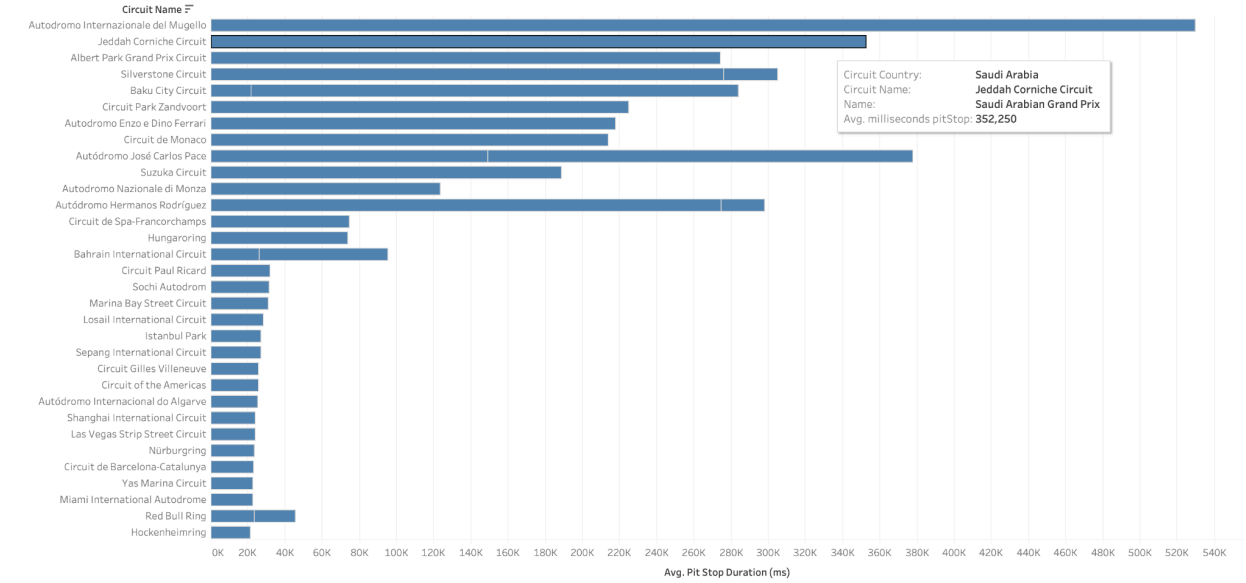
j) Which teams successfully adapt pit stop strategies mid-race, and what factors trigger these changes?

4) The data I have collected is the result of merging a couple different datasets relating to race results, circuits, drivers, constructors, and pit stops. The key columns, with their data type described, are listed below:
    - **id variables** (including resultId, raceId, driverId, constructorId): unique identifiers for each result, race, driver, constructor (nominal categorical)
    - **grid:** the starting position for a driver in given race (ordinal categorical)
    - **positionOrder:** finishing position order (ordinal categorical) – note, there is also positionText, which includes codes for drivers who did not finish explaining why they did not finish (these codes are nominal categorical)
    - **points:** points earned based on finishing position / fastest lap (quantitative)
    - **laps:** number of laps completed (quantitative)
    - **time:** race time in seconds and milliseconds (quantitative)
    - **year:** season of the race (ordinal categorical)
    - **round:** which race it is in the season (ordinal categorical)
    - **lat / lng:** latitude and longitude of the circuit (quantitative)
    - **circuitName:** name of the circuit (nominal categorical)
    - **constructorName:** name of the constructor/team (nominal categorical)
    - **forename / surname:** first and last name of driver (nominal categorical)
    - **stop:** which number pitstop in a given race (ordinal categorical)
    - **lap:** which lap that given pitstop was taken on (ordinal categorical)
    - **duration:** duration of pitstop in seconds and milliseconds (quantitative)

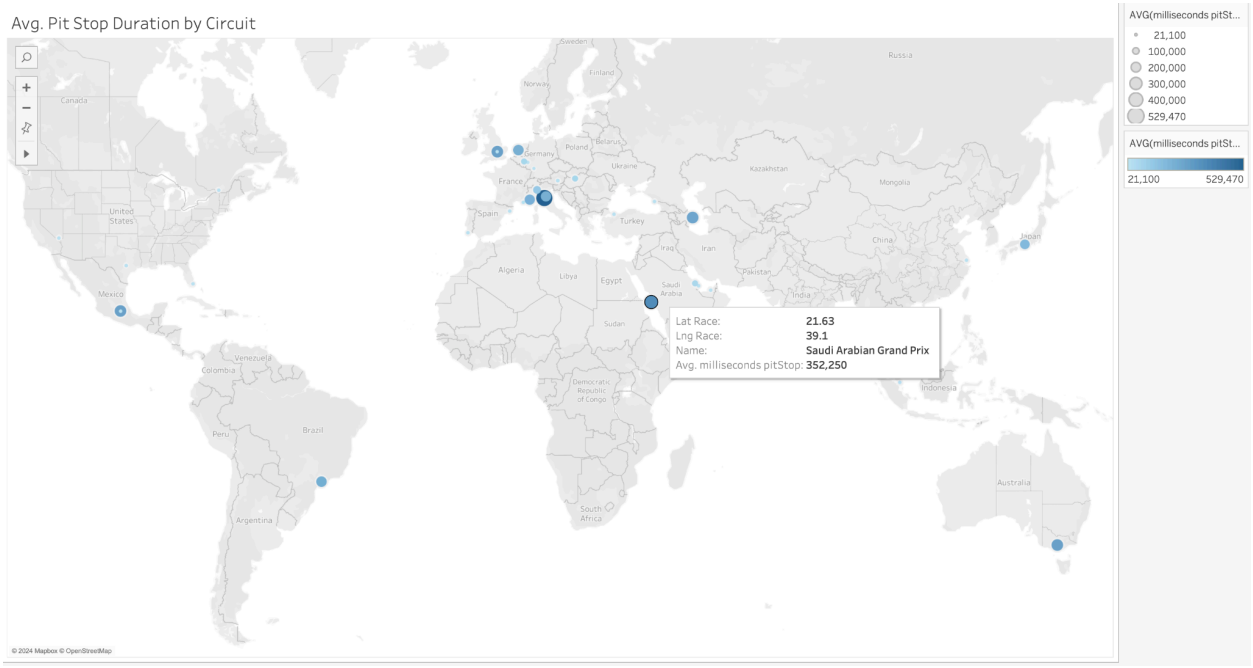
Tableau Work

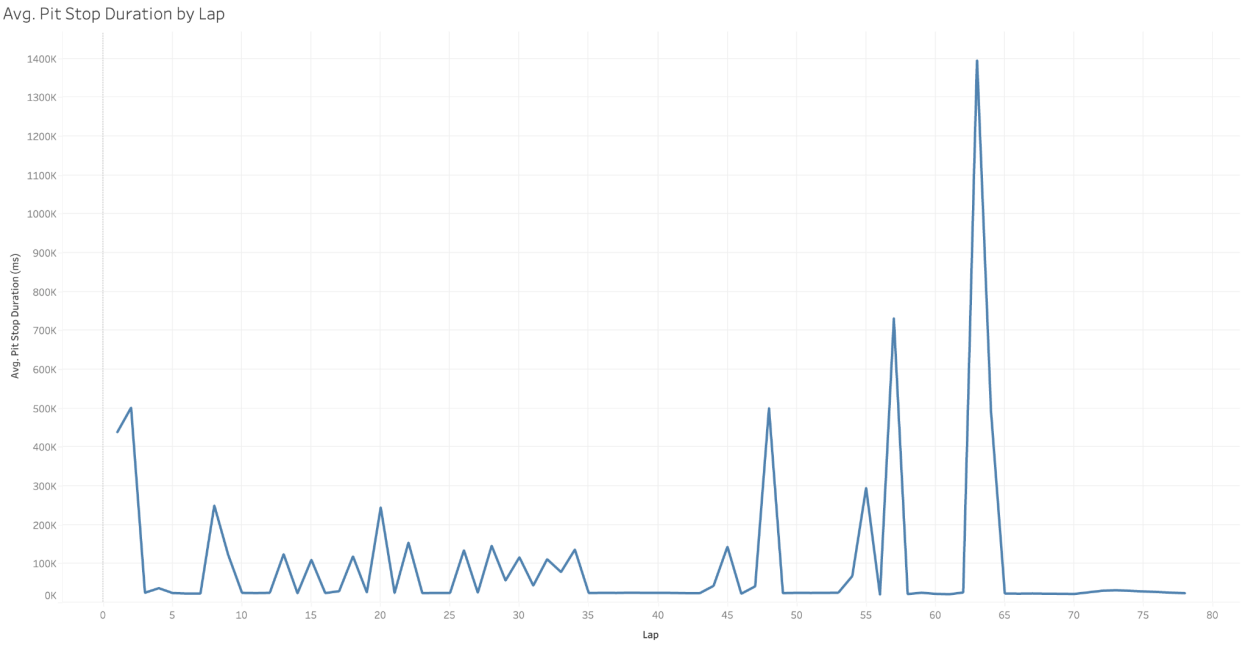
Screenshots of my Tableau visualizations are below. Originally, my questions focused on more complex, multivariable analyses, such as understanding pit stop strategy changes based on race phases or driver/car ability. However, in Tableau, it became evident that simpler, direct questions – like analyzing pit stop duration patterns by lap or constructor or circuit – yielded clearer insights. This shift happened because visualizations excel at showing initial trends and other elements of data exploration, but more complex questions involving numerous variables or conditional factors are often best addressed with more advanced data analysis, such as the analyses my group will be undertaking in CS1090a. Sticking to these refined questions allowed me to create effective and visually intuitive insights that will hopefully inspire the visualizations I will include in my web-app-based EDA.

Avg. Pit Stop Duration by Circuit



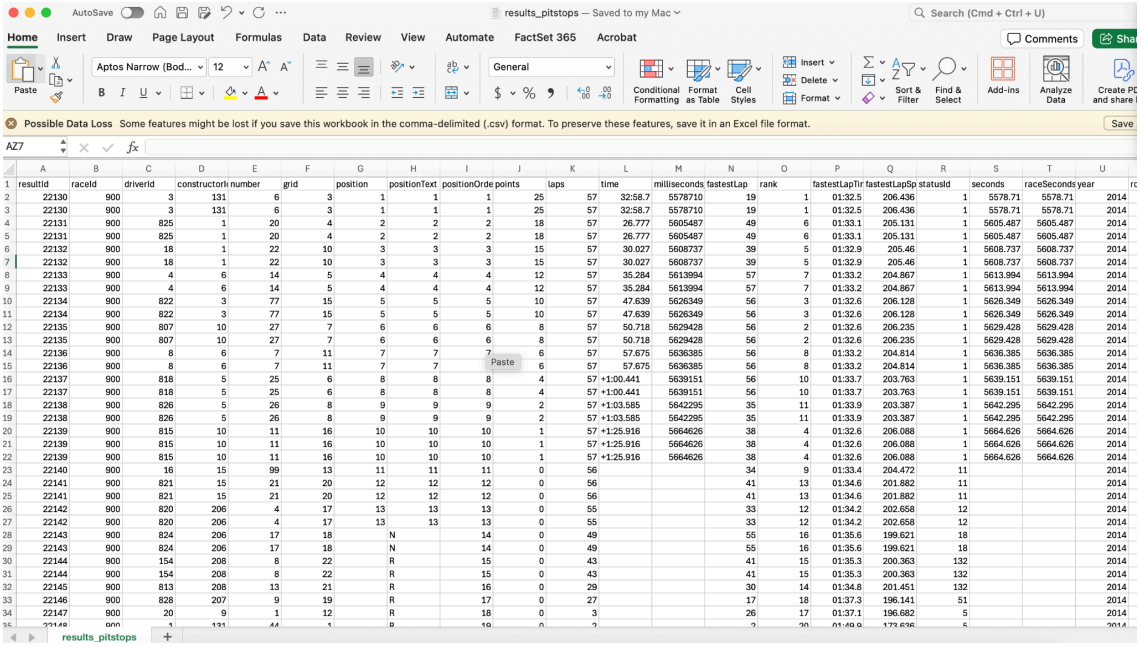
Avg. Pit Stop Duration by Circuit





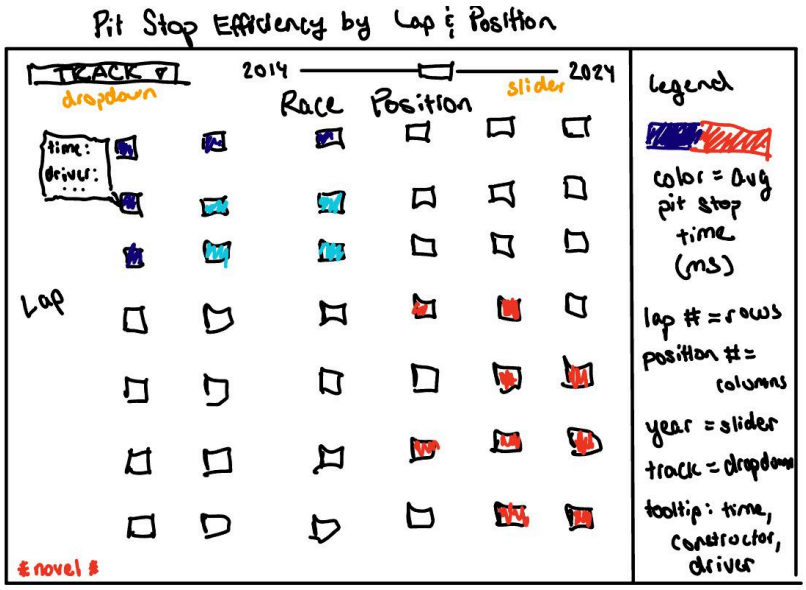
Final Project: Data, Sketches, Decide, and Storyboard

DATA: My dataset has been combined from multiple sources on the original Kaggle page (https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020). Below is a screenshot of the .csv file, with the data cleaned and combined, since I was not sure how else to submit the combined/cleaned data. If there is another way I should submit this dataset (Canvas will not allow me to upload it), please let me know.

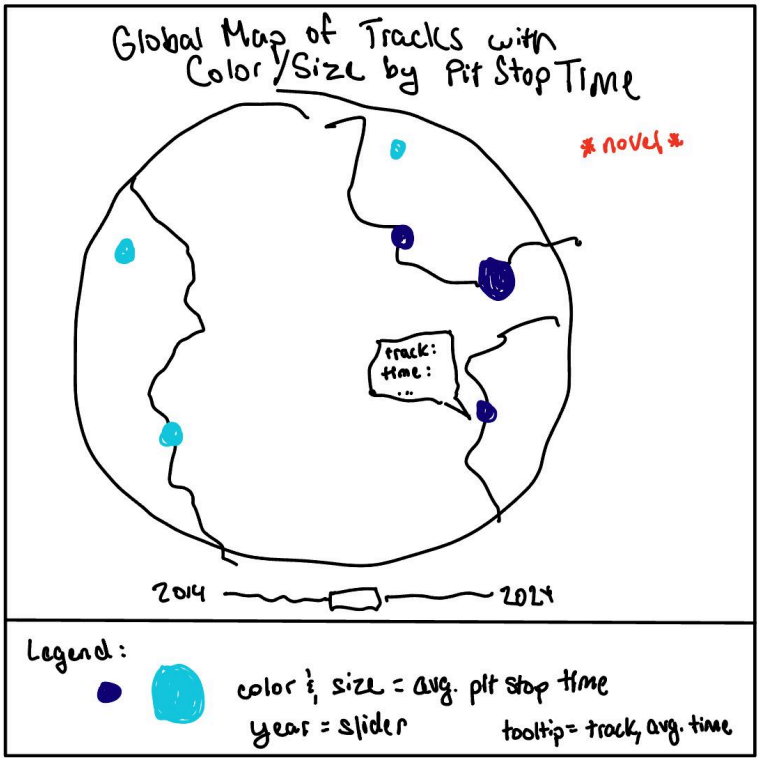


SKETCHES:

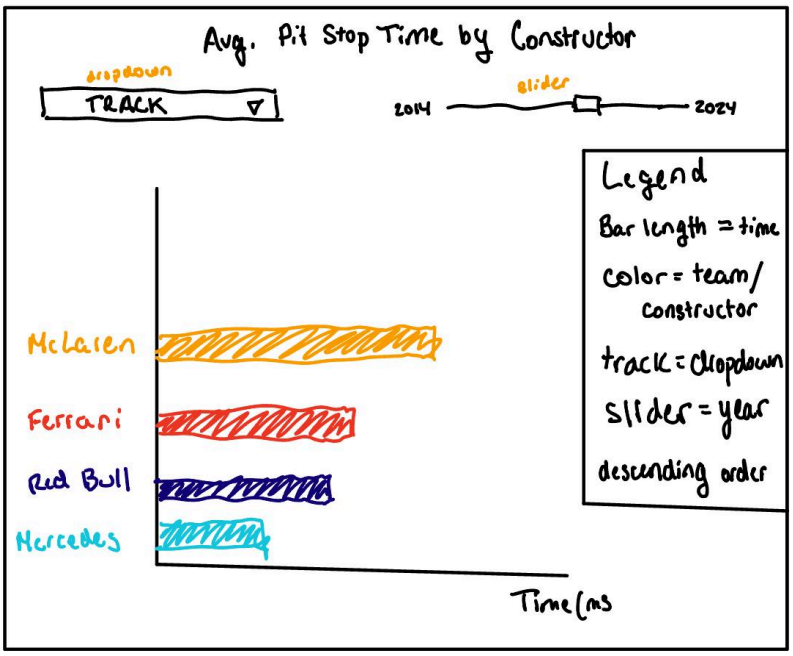
Question(s): What is the optimal pit stop lap for different circuits based on historical data? How do pit stop durations differ across circuits? Are there specific lap windows during which pit stops result in the most significant position gains?



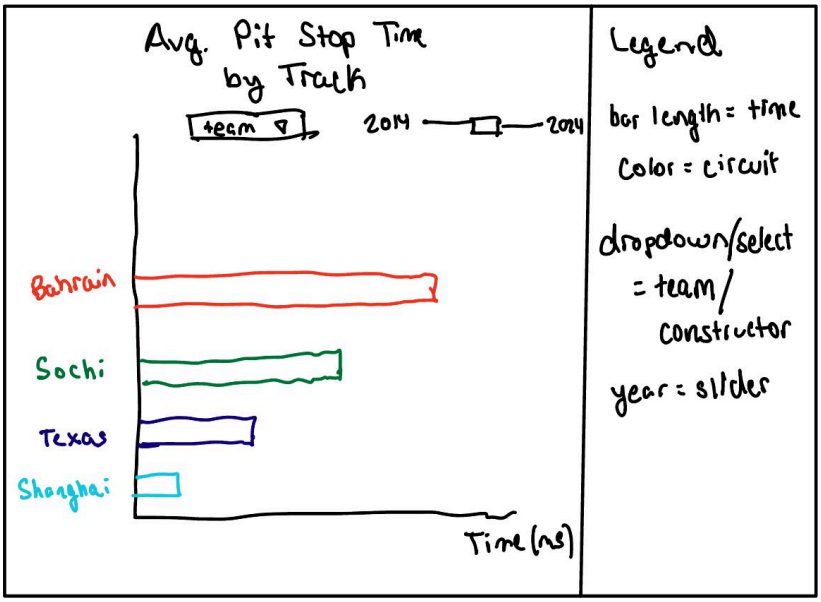
Question(s): How do pit stop durations differ across circuits?



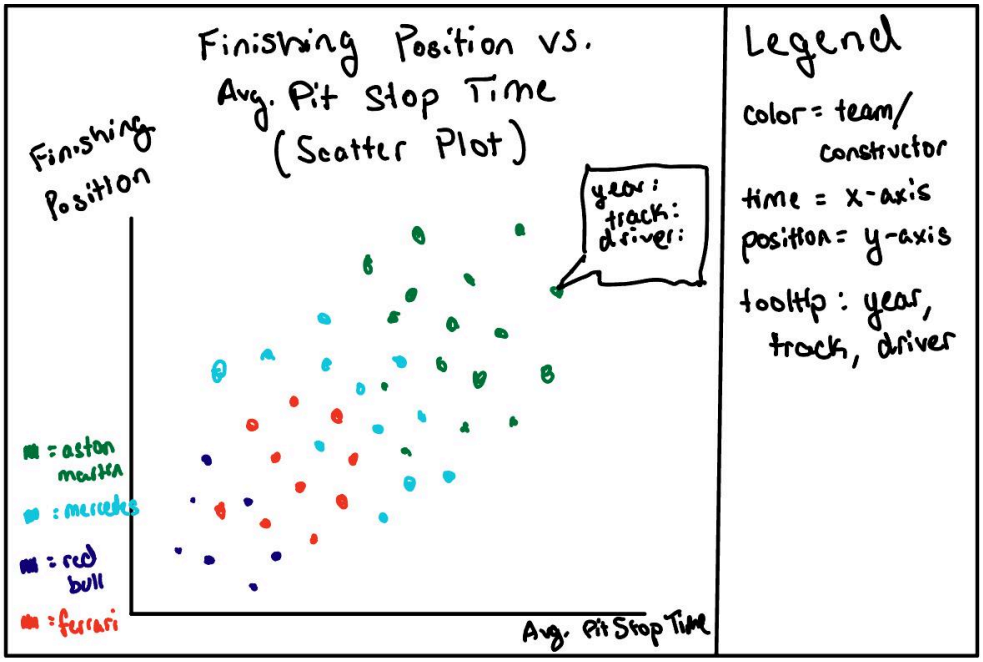
Question(s): Which teams demonstrate the highest pit stop consistency? Which teams successfully adapt pit stop strategies mid-race?



Question(s): How do pit stop durations differ across circuits?



Question(s): What is the relationship between the number of pit stops and final race positions for different circuits and race conditions? Which constructors achieve the best balance between pit stop speed and race performance?

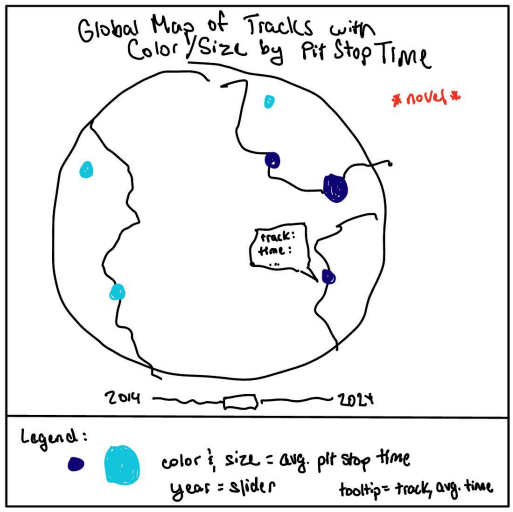


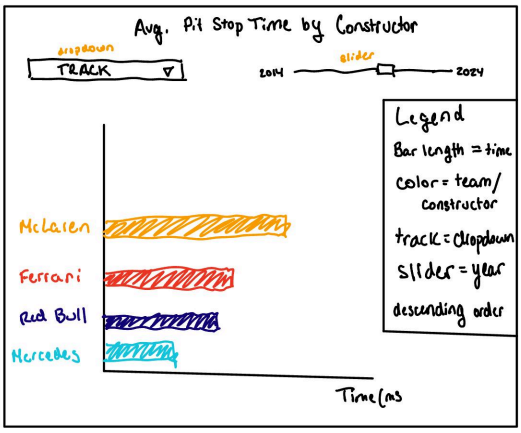
DECIDE:

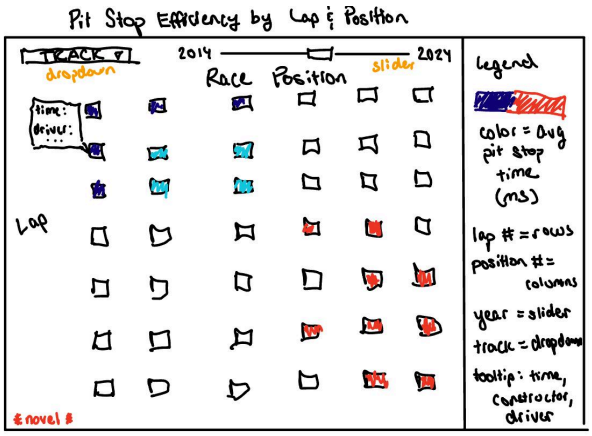
I create an affinity diagram below, but because I am the only member of my group, there is no voting process for which visualizations to create from those sketched above.

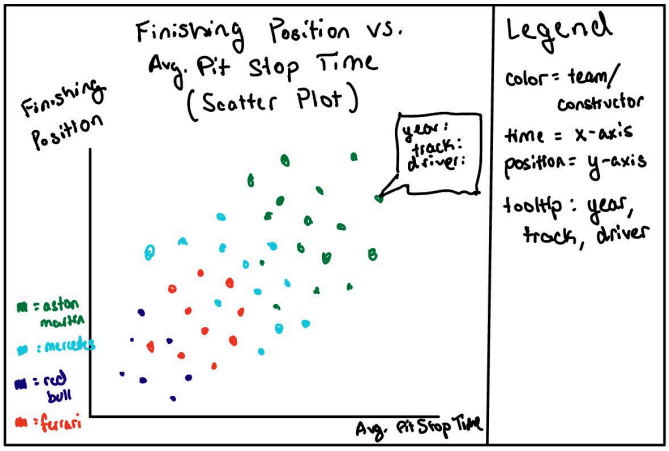
Sketch ID	Question ID	Author
1	1, 2, 5	JC
2, 4 (these show similar things in different ways)	2	JC
3	3, 10	JC
5	7, 8	JC

<u>Selected Sketches (in order from most relevant to least relevant):</u>


Global Map of Tracks with Color/Size by Pit Stop Time
* novel *
track:
time:
2014
2024
Legend:
color & size = avg. pit stop time
year = slider
tooltip = track, avg. time


Avg. Pit Stop Time by Constructor
dropdown
TRACK
slider
2014
2024
Legend
Bar length = time
color = team/ constructor
track = dropdown
slider = year
descending order
McLaren
Ferrari
Red Bull
Mercedes
Time (ms)


Pit Stop Efficiency by Lap & Position
TRACK
dropdown
2014
2024
slider
Race Position
time:
driver:
Lap
* novel *
legend
color = avg pit stop time (ms)
lap # = rows
position # = columns
year = slider
track = dropdown
tooltip: time, constructor, driver



Rationale for Decisions:

In selecting these sketches, I prioritized visualizations that maximize insights into pit stop strategies and performance comparisons across teams and circuits. The "Global Map of Tracks" was chosen as the most relevant view, as it provides an intuitive geographic overview of pit stop durations by circuit, helping users quickly identify which tracks have longer or shorter average pit stop times. The "Average Pit Stop Time by Constructor" visualization follows, focusing on team-specific performance; this bar chart enables a direct comparison of pit stop efficiency across constructors, with additional filters for year and circuit to explore changes over time. Next, the "Pit Stop Efficiency by Lap & Position" matrix offers a detailed look into how pit stop timing impacts race position and strategy, allowing users to identify optimal pit stop windows. Lastly, I included the "Finishing Position vs. Average Pit Stop Time Scatter Plot" to visually correlate overall race performance with pit stop efficiency, which may reveal whether shorter stops contribute significantly to better race outcomes and, again, how this relationship differs by constructor.

STORYBOARDING:

2-3 Insights from Exploring Data:

- There is a very wide amount of variance in average pit stop times for different tracks, weather this is because certain tracks degrade tires more, lead to cars being damaged more easily, or some other reason.
- There is also fairly wide variance in pit stops across different constructors/teams, although this seems to be less wide than by track

- There does not, however, seem to be a very strong correlation between lower pit stop duration and better finishing position — this is interesting, given how much time and effort constructors put into having efficient pit stops.

Main Message:
Combining these insights, my main message / insight is that:

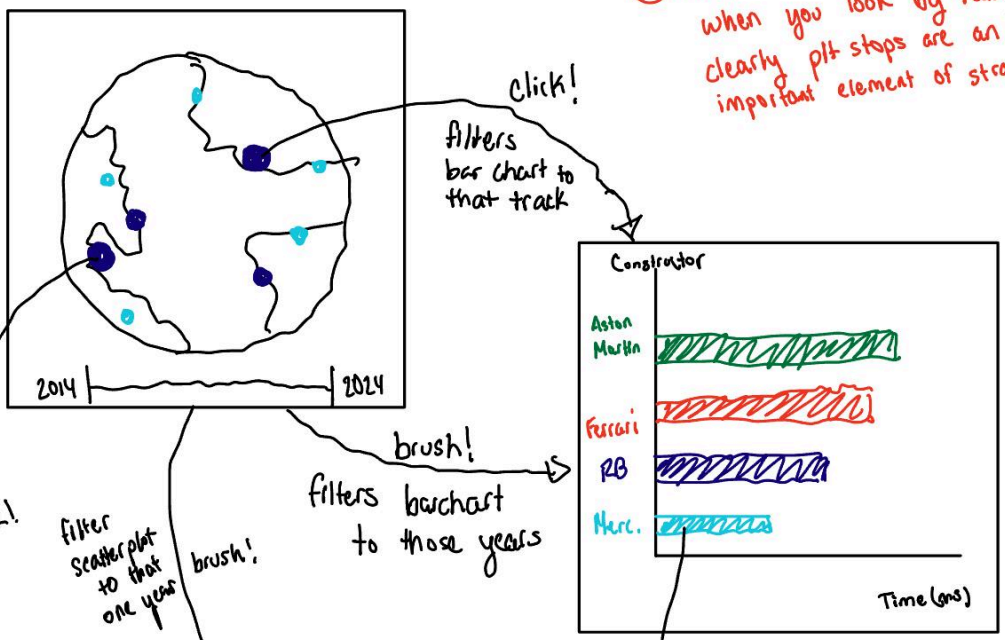
- There are very wide differences in pit stop times across both different teams and different circuits, but, despite these differences, there does not appear to be a very strong correlation between pit time and finishing position, suggesting that teams maybe spend too much time trying to become efficient in the pits and should allocate more resources towards other areas.

I chose this main message because I think it is interesting to see a predictor variable such as pit stop time that has so much variance across different teams and tracks but is nonetheless not that important for results. I would guess that part of the reason teams spend so much time trying to shave time off their pit stops is because they recognize this variance but have failed to see that these changes do not actually do much for results – I hope that displaying this message would be helpful for my audience of F1 racing/strategy analysts.

Storyboard:
See my storyboard on the following page.

(1) Hook: wide variation by track in pit stop time

(2) continued variation when you look by team, clearly pit stops are an important element of strategy



(3) main message: despite this variation, very insignificant relationship b/t pit stops and finishing order

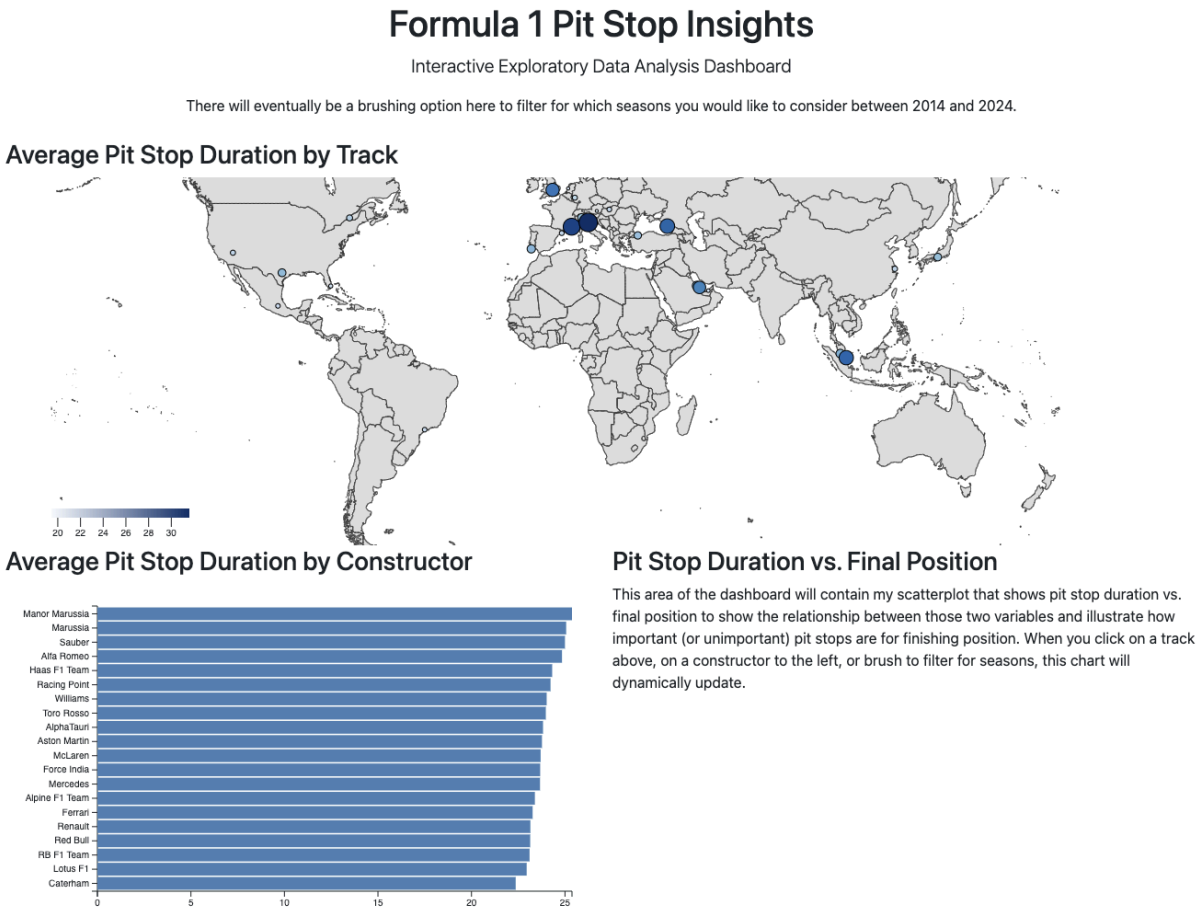
(4) solution: teams should devote less time/resources to reducing time in pits (maybe even compare pit stop duration with other predictors of performance to show this)

Final Project: Prototype V1

Name of student(s) who worked on prototype V1: Jameson Cohen

Data: Data downloading and cleaning is complete, and the data is entirely ready to use in my visualizations.

Prototype V1 Screenshot (zoomed out to see entire screenshot, but eventually dashboard will all fit on one page without scrolling):



Next Steps for V2 Prototype:

- Add the scatterplot visualization.
- Make it such that, when you click on a track on the map visualization, all the other data filters for that track.
- Make it such that, when you click on a constructor in the barchart, the scatterplot data filters only to that constructor.
- Add brushing functionality to allow users to filter the seasons between 2014 and 2024.
- Add background colors on theme with Formula One, and color the barchart in accordance with the team colors of each individual constructor.
- Make this dashboard fill the whole page without any scrolling necessary.

Notes: See zip file submission/index.html to see initial dashboard implementation. In the implementation, I include notes about my next steps for my V2 Prototype, most of which include linking the three views together through filtering by track (click on the individual track circles on the map), filtering by constructor (click on the individual constructor bars in the bar chart), and filtering by year/season (using brushing). I also am going to add colors to the barchart (based on team colors) and background colors for the entire web page in theme with Formula One. Finally, I plan to add the scatter plot visualization and, as mentioned above, make the whole dashboard fit in one view without scrolling.

Final Project: Prototype V2

Name of student(s) who worked on prototype V2: Jameson Cohen

Data: Included in data folder of zip folder upload on Canvas

Key updates since prototype V1:

- Implemented working scatterplot visualization
- Changed coloring of both barchart and scatterplot to match team colors for each individual constructor/team in F1, which would be more familiar to F1 data analysts
- Implemented **ALL** filtering and connected all three views, including:
    - Clicking on a particular track in the map visualization filters the data in the barchart and scatterplot to that particular track
    - Clicking on a particular constructor in the barchart further filters the scatterplot data to that particular constructor
    - Added a slider that allows the user to select which years between 2014 and 2024 (hybrid era of F1) they would like to view data from
- **All three views are now complete and working!**

Remaining steps for final submission:

Most of what remains for the final submission is styling, as well as a couple potential "reach goals" that are beyond the initial scope of what I set out to accomplish but that I would like to at least attempt to complete.

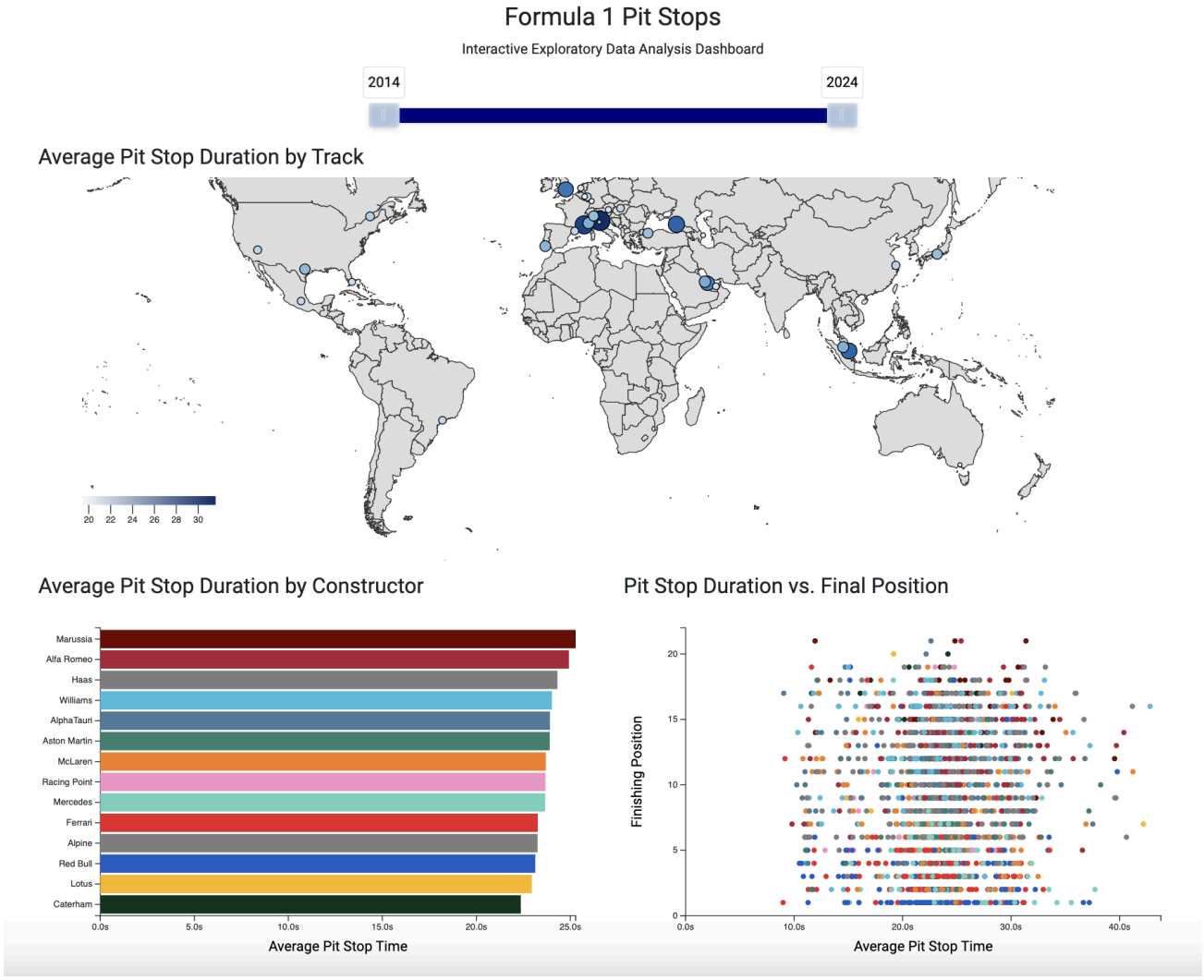
Styling steps include:

- Figuring out getting the entire dashboard to fit on one webpage with no scrolling required – I have been struggling to figure this out on my own in prototype V2, so this is something I may seek help with in office hours!
- Styling the entire dashboard a little more thoroughly, including background colors, text colors, etc. that would be more on-theme with the Formula One theme of this project

Reach goals include:

- Making it that you can select multiple tracks or constructors at once when filtering, as opposed to just one
- If my CS1090a group has completed enough of the finalized modeling steps in time, it would also be cool to create some sort of additional visualization of my team's model results, in order to compare insights from an initial EDA with those from a rigorous linear model. Note that this goal is somewhat dependent on how much my CS1090a group gets done, but I would at least like to try and accomplish the first of these reach goals if time does not permit me to accomplish this one.

Screenshot:


Formula 1 Pit Stops
Interactive Exploratory Data Analysis Dashboard
2014
2024
Average Pit Stop Duration by Track
20
22
24
26
28
30
Average Pit Stop Duration by Constructor
Marussia
Alfa Romeo
Haas
Williams
AlphaTauri
Aston Martin
McLaren
Racing Point
Mercedes
Ferrari
Alpine
Red Bull
Lotus
Caterham
0.0s
5.0s
10.0s
15.0s
20.0s
25.0s
Average Pit Stop Time
Pit Stop Duration vs. Final Position
Finishing Position
0
5
10
15
20
0.0s
10.0s
20.0s
30.0s
40.0s
Average Pit Stop Time

Final Project: Testing / Think-Aloud Study

Notes:

- I was not able to be in class in-person for testing because I had already traveled home for Thanksgiving, so Saba Mehrzad conducted my think-aloud study over Zoom
- Per the instructions for Week 13 - Test, since my group consists of just one person (less than 3 people), I only conducted one think-aloud study, with results shown below.

	Notes (To be filled by project leads)
Tester Name	Saba Mehrzad (sabamehrzad@college.harvard.edu)
Describe any usability issues or confusion the tester encountered while using the prototype.	- Confusion about where to click on tracks to filter; the user mistakenly clicked on the instruction text instead of the circle. - Trouble unfiltering tracks initially. - Scatterplot malfunctioned when filtering by track, constructor, and then unfiltering both (not all points reappeared).
Was the tester able to understand the main message of the data story? (e.g., Yes/No + why/why not?)	Yes. The tester noticed several key patterns, such as the performance consistency of powerhouse constructors/teams such as Red Bull, Ferrari, Mercedes and the lack of a strong relationship between pit stop time and finishing position, even after filtering. The tester quickly picked up on the data story, namely that there is a fair amount of variation in pit stop time by track and constructor (which remains when you filter by seasons, individual teams, etc.), but pit stop time does not seem to have a very strong relationship with finishing position on its own (you probably need to control for other variables).
What parts of the interface or visualization did the tester find most engaging or effective?	- The interactive slider for filtering seasons and the hover functionality on the bar chart and scatter plot - Being able to filter by constructor and observe team performance trends in the resulting filtered scatter plot over time
What parts did the tester find confusing or less effective?	- The variation by track in the world map did not immediately stand out to the tester – they suggested a more clear legend showing indicating the size and color variation in pit stop duration by track

Did the tester encounter any inconsistencies in design, data, or narrative?	Not really. The tester had some experience with F1, and they noted that the app uses current team names for teams whose names have changed over time, which initially confused them but they understood quickly afterwards.
Were there any unexpected interactions or insights that emerged during the session?	- The tester raised insightful questions about what contributes to pit stop durations (e.g., routine tire changes vs. repairs). - They identified that constructors perfecting pit stop timing correlates with better and more consistent finishing positions, which could inspire further modeling. - These insights are not exactly unexpected but actually exactly what we want to see from an exploratory data analysis (EDA) – we want the user to be asking all the questions that they would eventually want to explore with more robust modeling / data science techniques (which is exactly what my team is doing for the CS1090a project).
What specific improvements or changes did the tester suggest for the prototype?	- Make the instructions for how to filter more clear (so user does not click on the "click to filter" text but on the actual track circle) - Clearly highlight or annotate the variation by track on the world map - Fix filtering and unfiltering bug that occurred with colors on bar chart
Did the tester suggest any additional insights or visualizations to include?	The only other insight the tester expressed interest in seeing was the factors influencing varying pit stop durations, such as whether it is a routine tire change or if the team has to make a larger repair (e.g. to the front wing of the car).
General observations or comments from the tester.	The tester seemed to really enjoy interacting with the F1 EDA and gaining an initial understanding of the data. They provided valuable feedback for improvement, such as making filtering instructions more clear and more vividly highlighting variation by track. It was really nice to see them interacting with the web-application exactly how we would want someone to interact with an EDA – gaining initial insights about the patterns in the data and asking interesting questions that one would be able to explore with more robust modeling techniques (the exact purpose of an EDA!).

“Post-Mortem” Questions from Think-Aloud Study:

Based on the results of your ‘think aloud’ study, what would you improve in your data story?

Based on the results of the “think aloud” study, there are two key things I would improve in my data story to make it more accessible to users:

- Firstly, I would make track filtering more intuitive for users, so that they understand immediately where to click in order to filter the data for particular tracks as opposed to struggling at first like the tester did in the study
- Secondly, I would highlight the track-level variation more clearly, potentially with a more clear color legend and a size legend as well, showing that different tracks vary quite a bit in the average pit stop duration that occurs in races at that track

In addition to these two key changes, I need to be sure to fix any filtering bugs that might occur, particularly the one I noticed with the colors on the bar chart after filtering and unfiltering a couple times. I could potentially also provide some more context for team name changes, but I think that might confuse users more, given the teams are not actually changing, just their names.

Are there any additional insights and visualizations you would use? Would you amplify or change your message? Did your narrative work? Did the tester get your takeaways?

I do not know that there are any new insights or visualizations I would add to improve my narrative. The tester quickly understood the data story I was trying to tell with my interactive EDA. The only thing I would change, as mentioned in the question above, is making the initial insight that track variation is quite wide a little more clear by using better legends and more clearly highlighting track-level variation.

Decide as a team which of these improvements you will implement and write down your decisions and why you made them in your process book as a numbered list.

1. Make filtering instructions, particularly for track filtering, more clear.
    a. MUST-HAVE
2. Improve the legend and add any necessary annotations for the world map to more clearly highlight track variation in pit stop duration.
    a. MUST-HAVE
3. Test for filtering and unfiltering bugs in all visualizations and fix them.
    a. MUST-HAVE
4. Include some explanation for former team name changes and the fact that all team names in the visualization reflect the team’s current name as of 2024.
    a. OPTIONAL

Implement the intended changes and check them off your list (e.g., adding "done"). You can distribute the tasks among your team members. If you are unable to implement specific changes, please explain why and describe the expected results in your process book.

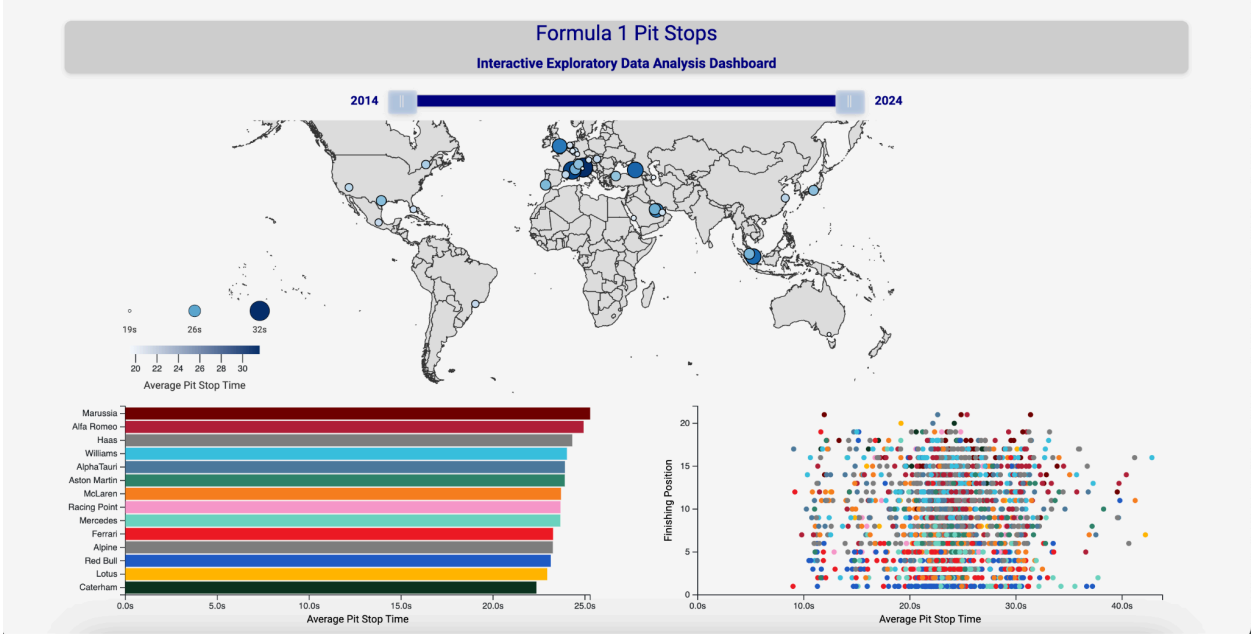
I am implementing these changes for my final submission, now that my "think-aloud" study is complete, and will check off each task in this process book as I complete it. I am working alone on this project, so all of these tasks will be completed by me as opposed to by any team members.

Final Submission Notes / Checklist

Feedback Implementation Checklist:

- ☑ ~~Make filtering instructions, particularly for track filtering, more clear~~
- ☑ ~~Create size legend in globeVis to make track variation more clear to users~~
- ☑ ~~Test for filtering and unfiltering bugs and fix them~~
    - NOTE: The filtering issue is when you filter a track and then unfilter it, the scatter plot does not unfilter
- ☐ ~~OPTIONAL: include explanation for former team name changes~~
    - I chose not to make this "OPTIONAL" change, after thinking carefully about it. This is because my selected audience is F1 analysts and professional strategists. Given their background in the sport, they would be familiar with the most updated team name changes and understand that, in earlier years, those teams would technically have had different names. Given this, I actually thought that adding a note / explanation explaining this would unnecessarily draw their attention away from the key points of my data story (variation by track and constructor in pit stop duration, lack of clear relationship with finishing position without other controls, etc.)
- ☑ ~~My own feedback: Make dashboard take up one page without scrolling~~
- ☑ ~~My own feedback: Add some background colors and styling to the title, which I did not get to do previously, to make the page look nicer~~

Below is a screenshot now that I have implemented all the feedback (both from my testing and my own feedback):



Additional Checklist for Final Submission:

- ☑ ~~Create readme file on how to run code, used libraries, etc.~~
- ☑ ~~Prepare data for submission and create data description document~~
- ☑ ~~Host website on Github pages and prepare usable link~~
- ☑ ~~Create 2-minute screencast~~
- ☑ ~~Fill out peer evaluation form~~

Link to Final Project Website: jamesoncohen02.github.io

Link to Final Screencast:
https://drive.google.com/file/d/15l8xqVwZL2_1HUIjA0enlj8uMTmsEDVe/view?usp=sharing