Web Scraping and Social Media Scraping - Final Project Description

Names and ID's of all participants.
Vladimir Shargin (437981)
Vadym Dudarenko (444820)
Ivan Grakhovski (444422)

Short description of the topic and the web page.
Formula 1 is the highest class of international racing for open-wheel single-seater formula racing cars sanctioned by the Fédération Internationale de l'Automobile (FIA). The World Drivers' Championship, which became the FIA Formula One World Championship in 1981, has been one of the premier forms of racing around the world since its inaugural season in 1950.

The goal of the project is to scrape results for Formula 1 races for all of its history starting from the year 1950 until now. The data to be scraped includes basic information about the race itself (season, location, full name, date, track name), drivers' finishing positions, team names, number of points gained, number of laps completed and time at finish.

Formula1.com is the official website for Formula 1, the starting page to be used in this project is the results archive: https://www.formula1.com/en/results/archive-1950-2016.html

Short description of your scraper mechanics - what the program is technically doing.
Selenium-based scraper:

- Fetching the starting pages for each season
- Scraping the links to every race result for the season
- Fetching the pages with race results
- Scraping the result tables and saving information to a list of dictionaries
- Converting the list to a pandas dataframe and exporting the results to a CSV file

Scrapy:

- Extracting links for each season
- Scraping places to every year on the current website
- Extracting results, cars, participants and other information for all links
- Saving the results to a CSV file

BeautifulSoup:

- Getting the links for every year in this website
- Scraping countries and places to each season
- Getting data about races
- Saving the results to a CSV file

Short technical description of the output you get.
The output is a CSV file with the following columns: season, race name, full race name, race date, race location (track), driver position at finish, driver's name, driver's car (team name + engine manufacturer), number of laps completed, time at finish, points gained. Each row represents a single driver participating in a single race.

Extremely elementary data analysis - you need to prove, that collected data can be used for further analysis, but nothing more (hard limit of data analysis: one page).
The amount of data collected is 1062 races over 72 years.

The driver who has the most wins to date (May 14th 2022) is Lewis Hamilton, with 103 races won. Next in the top 5 are: Michael Schumacher (91 wins), Sebastian Vettel (53 wins), Alain Prost (51 wins) and Ayrton Senna (41 wins). All of these drivers have been awarded the World Driver Champion title at least once in their career.

As of the time of the writing, in all of Formula 1 history out of 770 drivers who have started in at least 1 Grand Prix, 34 drivers won the World Drivers' Championship at least once, with Michael Schumacher and Lewis Hamilton winning it 7 times.

Sebastian Vettel and Michael Schumacher hold the record for most wins during a season, 13 with Vettel achieving this in 2013 and Schumacher in 2004.

British driver Stirling Moss has won 16 races in his Formula 1 career, but never won the WDC, holding the highest number of wins without a WDC.

The data can be enriched with other variables, for example qualifying positions (from Formula 1 website) weather (which can also be scraped from a weather archive website), and for model building feature engineering can be applied to add lagged features, construct features based on performance to date, reliability based on number of DNFs, etc.

As seen in academic papers, the information can indeed be used for further statistical analysis of how drivers' skill, cars' performance and other factors affect Formula 1 results, as evidenced in a paper by Bell, A., Smith, J., Sabel, C. E., & Jones, K. (2016) "Formula for success: Multilevel modelling of Formula One Driver and Constructor performance, 1950–2014" (https://eprints.whiterose.ac.uk/96995/). The authors apply a multilevel mixed effects model to analyse drivers' performance and see what factors influence the results the most.

Detailed description which group participant wrote which part of the project.
We have divided the project work so everyone worked on one of the three approaches, but everyone is familiar with all the presented tools. Short description and analysis is shared work.

BeautifulSoup - Ivan Grakhovski
Scrapy - Vadym Dudarenko
Selenium - Vladimir Shargin