Historical code: https://github.com/wsheehan19/Equibase

1. Gather Data
    a. Go to https://www.equibase.com/ & Navigate to Results/Full Charts
    b. Click on any track, any day and click View the Full Card Here
        i. example: https://www.equibase.com/static/chart/pdf/DMR082522USA.pdf
        ii. Example 2: https://www.equibase.com/content/rotd/fullpp.pdf
    c. We need to gather, save & process as many historical race cards across as many tracks as possible in an automated fashion.
2. Extract & Load Data (Data Cleaning)
    a. Given a PDF, how do you extract information from this into a machine readable format. The format is needed to turn the data into a structured, schematized, tabular format. The tools used will be known as OCR or optical character recognition tools.
    b. Data will need to be loaded into a database, so a schema will need to be created based on the early extractions to understand how a horse or a race are going to be connected together.
        i. PostGRESQL or MySQL are the most common DBs for these applications and are relatively cheap and easy to use
    c. Once the data for a few dozen races is extracted and loaded, we can begin to build our data transformations.
3. Transform Data (Data Cleaning)
    a. This is the data modeling component, where data from each individual race will be turned into a usable format. The exercises here will include basic transforms like one-hot encoding and normalization, or more complex transforms, like nesting race splits together to make it normalized and usable
4. Exploratory Data Analysis
    a. After we have an easy way to identify winners and have begun to create features from our winners we will need to conduct some basic analysis to validate known hypotheses. These questions will get us started as examples for how we might structure our data in our ELT process and how we might begin to think about building or deploying any form of model.
        i. What is the distribution of winners by starting post?
        ii. What is the distribution of races by track?
        iii. What is the distribution of races by track & track condition?
5. Build ML & Prediction
    a. Once we have a structured tabular dataset that we can evaluate, we can build both supervised and unsupervised machine learning models to determine the winner of a race and identify characteristics of future winners. There are a number of techniques we will apply here as we test, but before we get here we will need to have a sufficient number of records that are cleaned and transformed.
6. Other future steps:
    a. Analyze and Validate predictions

    b. Deploy & Test Model
    c. Iterate pipelines & improve automation of processes
    d. Turn into application

What can be done in parallel?

1. Data ELT & data gathering can be done in parallel as long as the communication between the two tools is async.
    a. Note: manually collect PDFs to begin data ELT
2. Once data has been transformed, analysis code can be written while data collection & processing is still ongoing
3. Data transformation is an iterative process, we should expect that we will need to iterate here based on the results on analysis and data collection
4. Model building code can be written in parallel to data analysis, but will not be accurate until the model is analyzed, tested, and sufficient data is provided
5. Process optimization and code improvements can always be added along the way to support modeling code

Expectations by week (accelerated):

1. Historical Code review, project planning, initial distribution of duties
2. Data Collection & ELT
3. Data Collection & ELT
4. Data Collection & ELT
5. Data Collection & ELT
6. Data Transformations
7. Data Transformations
8. EDA
9. EDA
10. EDA
11. Modeling & Prediction
12. Final Write Up & Documentation

Alternatively, there is a world where the entire semester is spent on data collection and ELT. Progress is progress for Randy and I, but I want to ensure you as students get the most out of these 12 weeks.