Step 1: Understand the Instruction Video

Two types of racing: harness and flat racing

Trot racing walk/the (slowest gait), gallop (fastest gait), Trot (move two legs that are diagonal forward at the same time)

Trot - **Mounted racing** (when the jockey rides his horse on a saddle) and **Harness racing** (when the jockey rides on a two-wheeled cart)

The horse pulls the driver on a two-wheeled cart called a sulky

Jockeys called drivers during harness racing

The same horse is allowed to race both ways harness or mounted racing
"Heroes" of Trot racing
Trotting horses are the best athletes
In France, racing horses are allowed to participate until they are 10 years old
A great champion is called a "CRACK"
The horse's Owner deals with a Trainer to get his horse in racing shape
The horse goes to training everyday in the Trainer's stable
The Trainer establishes with his team a training program and chooses which races to attend to (but the driver/jokey is the one who will race with the horse)
The driver wears the Owner's color jacket called Casaque

Note: Volt start (let the sulkies move in a specific area) vs Auto start (led by a moving car with a starting gate, horses follow, then the car speeds up and the race starts)
To win, you need to finish in first place without galloping.

How are racing organized?
Important: drier's tactics during the race VARY

The better horse, the higher division. Best horses race in the "Groupe 1" Division

Step 2: Convert Every Variable into Everyday Language

Colour Codes
Orange highlight: VERY important
*Non-significant predictors are all ~~crossed out~~
Italicized : reconsider
Yellow: needs further discussion

— Sean —

~~**Age Restriction**~~

~~Age restriction of the competition~~
~~Only horses with certain ages are allowed to enter this competition~~

Barrier

Starting Position - Volte Start gate position

BeatenMargin

How far the horse was behind the winner in that tournament
0 would mean the winner of the race

~~***ClassRestriction***~~

~~*Restriction of a class in the competition. Only horses with a certain class can join the competition.*~~

~~CourseIndicator~~

~~Horse racing track surfaces~~
~~Due to MISSING VALUES let's omit this variable~~

DamID

ID of a mother of the horse

Disqualified

Determines whether the horse got disqualified in this tournament.

Distance

Full distance of the racing course at the tournament

FinishPosition

Winning position of the horse in the tournament. 1 indicates a winner.

~~***FoalingCountry***~~

~~*Country where the horse was born*~~

~~**FoalingDate**~~

~~Date when the horse was born~~

FrontShoes

Refers to the horseshoes that are placed on the front hooves of a racehorse; basically metal shoes

(either front legs only or behind legs only)

Gender

Gender of a horse
matters especially for gender specific case

~~**GoingAbbrev**~~

~~Refers to abbreviation to describe the condition of the race course. It explains the condition of a racetrack in terms of its firmness or softness, and it can greatly impact a horse's performance.~~

~~GoingID~~

~~use either goingID or going abbrev (don't include both because they are supposed to be "GoingID" has a corresponding numerical code by "GoingAbbrev"~~

— Marvin —

~~HandicapDistance~~
~~Horses with higher earnings have more distance to cover, this is additional distance. It occur only in 0, 25 meters and 50 meters~~
~~**Why does it have to have a negative value?~~
~~**Why is this the same for each tournament?"~~

~~HandicapType~~
~~This occurs as blank or cwt or hcp, i cannot find its meaning in guide book, and it appears to have no correlation with handicap distance.~~
~~Racetype and handicap type are grouped them together and handicap distance~~

HindShoes
This is to show if the horse is wearing shoes on its hind legs. If it's not wearing its 0, it's wearing its 1 => it can be just 0, 1, 2, 3
HorseAge
The age of the horse

HorseID
The id of the horse

JockeyID
The id of the jockey, who is the horse rider. Note this only occurs when its mounted race.

PIRPosition
PIR is a way to evaluate how well a horse has adapted to its new racing environment and the level of competition it is facing after being imported. It's often used to assess a horse's performance in its most recent starts.
*It's numbered from 0- 26(max). *(I think) 0 indicates it's a volt start game, so there is no position.*

PriceSP
Lower the better (we can observe the trend)

Prizemoney

The amount of money the horse won in the race. Note only top 7 get prizes, and it's a proportion of total prize money.

RaceGroup

Seems relevant

Race groups (class levels) to indicate the quality and competitiveness of the race

Group 1, Group 2, Group 3

Ensure that horses of similar abilities compete against each other, which creates a fair and competitive environment.

~~***RaceID***~~

~~Id of a race~~

~~***RaceOverallTime***~~

~~The total amount of time it takes for a horse to complete~~

~~***RacePrizemoney***~~

~~Total amount of money as prize in the race.~~

~~***RaceStartTime***~~

~~The start time of the race, format is in (year-month-day hr:min:sec)~~

— Serena —

RacingSubType

T = trotting: standard trotting (vehicle is the standard)

T M = trotting monte: when the driver rides the horse rather than the buggy (buggy is just horse drawn vehicle)

Saddlecloth (in my intuitive sense it will most likely won't affect but we can consider if this affects FinishPosition or not)

Saddle cloth and finish condition

Equipment placed between the horse's back and the saddle: usually consider the following: protection, comfort, moisture absorption, sizing and fit. Style and personalization. Sponsorship and identification

Update: we figured saddlecloth actually do matter

SexRestriction

Gender restriction of a tournament between male and female horses.

In horse racing, "C&G" is an abbreviation that stands for "Colts and Geldings." It is often used to classify and describe a group of male horses in a race or in the context of horse racing. Here's what the abbreviation represents

Colts: young male horses, typically age below 4, that have not been gelded (castrated). They are still capable of breeding.

Geldings: Geldings are **male horses** that **have been castrated**. Castration is a common procedure in horse racing to reduce aggressive behavior and to allow the horse to focus on racing rather than breeding.
“C&G” means it is exclusive for colts and geldings, and fillies (female horses) are not allowed to compete in that particular race. The classification is based on gender and is used to create fair conditions for male horses.
Quick qs: is it C AND G or C or G? => this will be done with Rstudio for an analysis

SireID
ID of the sire/father

StartType
Startype of the race tournament. Eg: auto or volte
M and V
Volt start (let the Sulkies move in a specific area) vs Auto start (led by a moving car with a starting gate, horses follow, then the car speeds up and the race starts)

StartingLine
Where the competition starts. Starting point for the race to ensure a fair/organized start.

Surface
The type of ground/track on which the race takes place.
It is supposed to have a crucial role in determining race conditions, race strategies, and the performance of horses.

- Grass (or turf)
- Sand
- Ash and cinder (or similar, such as clinker or pozzolan).

TrackID
The track where the tournament was held.

TrainerID
ID of a trainer

NoFrontCover
- “Cover” In the context of horse racing, refers to a horse's position during a race. "Cover" can also be associated with the concept of trip handicapping, which is the assessment of a horse's race based on the path it took during the race.
- How is “cover” often used in horse racing?
Position in the pack: “cover” describes a position in relation to other horses during a race. Having a “good cover” represents when there’s other horses right ahead (so racing

behind/alongside with other horses). This is expected to provide a much more favorable position and conserve energy.
Trip Handicapping: Trip handicappers assess a horse's performance based on the path it took during a race. A horse that had a clean trip with good cover may have had an advantage over horses that had to race wide (away from the rail) or encounter traffic.

PositionInRunning
(i) easy position (ii) mid-race position (iii) late position (iv) finishing position
Understanding a horse's position in running provides valuable information about its performance. Race strategy, and the dynamics of the race.

WideOffRail
Horse's position on the racetrack relative to the inner rail, which marks the inside boundary of the track.

WeightCarried
The amount of weight which horse carried includes the weight of the jockey, saddle, and any additional weights assigned to the horse. In handicap races, horses are assigned weights by their past performances. For instance, better performing horses are oftentimes assigned higher weights while less performed horses are assigned to carry lighter weights.

WetnessScale
Refers to track conditions or the state of the racing surface. Usually weather related and maintenance. Information is provided to trainers, jockeys, and bettors such that they can help assess how the track might affect the race.
1, 3, 4, 7, 9? => let's double check

Step 3: Find out Correlation

Sean

- Short distance -> try to find out what type of horses are advantageous DONE!
- Age restriction -> do older horses perform better in general because they are better experienced? DONE!
- Class Restriction -> for a given class is there some correlation such that certain horses with certain features are more advantageous - what is exactly a class in our context? There are approximately 2200 classes…
- GoingAbbrev -> how do we know which horses perform better in a bad condition? (maybe age?)
- Handicap Distance -> check how many horses win with handicap vs without (why does it have negative values?, also why is this the same for each tournament?)

Marvin

- Figure out Handicap Type <- what do they really mean?

- An empty string ("") likely indicates that there is no handicap applied to a horse
- "Hcp" could stand for "Handicap" refers to any conditions that are used to even the chances of winning for different horses, which could include extra distance to be covered or additional weight to be carried.
- "Cwt" might be shorthand for "Catch Weight", where horses carry a weight determined by agreement between the owners, without reference to the official handicap.
- "SW" could represent "Set Weights", where horses carry a predetermined weight based on the conditions of the race, rather than a weight influenced by individual handicapping.

- Front Shoes & Hind Shoes why 0 to 3?

  Note, Only three possible combinations are allowed for the use of horse shoes in any given race: front legs only, hind legs only, or completely unshod.

- "0" - The horse is racing unshod on the respective legs.
- "1" - The horse has shoes on the front legs only (if FrontShoes is "1") or hind legs only (if HindShoes is "1").
- "2" - The horse has shoes on both front and hind legs.
- "3" - This value is uncertain, but it could possibly represent an shoeing condition not mentioned in the document

- *PIRPosition -> needs investigation*

- The PIRPosition likely stands for "Position In Running" and represents the position of a horse at certain points during the race.

- "0" could mean that the horse did not participate in the race or did not finish, possibly due to disqualification.
- The numbers "1" to "26" likely represent the running position of the horse during the race or at a specific point in the race, with "1" being the lead position and "26" being the last.

- Finish Position -> needs investigation for acronyms

- The alphabetic values correspond to various reasons a horse did not finish in a traditional ranked position, either due to performance issues, disqualification, or other incidents during the race.

- Numeric values ("1", "2", "3", etc.) indicate the actual finishing position of a horse
- "BS" stands for Break Stride, which means the horse broke its trotting stride, which can lead to disqualification in trot racing.
- "PU" stands for Pulled Up, meaning the horse was withdrawn from the race before it finished, often due to injury or not being competitive.
- "FL" indicates the horse Fell during the race.
- "NP" means Took no Part, suggesting the horse entered but did not start the race.

- "DQ" stands for Disqualified, meaning the horse was disqualified for some infraction during the race, like breaking stride as mentioned in the document.
- "UN" is not a standard racing abbreviation, but it could potentially stand for unplaced, indicating the horse finished but not in a position that earns prize money or points.
- "UR" is not detailed in the document, but in racing terminology, it often means Unseated Rider, where the jockey has fallen off the horse or been unseated.
- "WC" could mean Wrong Course, suggesting the horse took the wrong path or did not complete the correct course, though this is speculative as it is not defined in the document.

- PriceSP -> study the trend with finish position

- "SP" typically stands for "Starting Price." The Starting Price is the odds available on a horse at the start of a race.

- The numerical values given for PriceSP are likely to indicate the odds for each horse at the start of the race. These odds reflect the betting market's view of each horse's chances of winning. The lower the number, the more favored the horse is considered to be, and vice versa. For instance, a horse with a PriceSP of "2" is expected to have a higher chance of winning compared to one with a PriceSP of "26", hence it would offer a lower payout because it's more likely to win.

Using price, race group related data to discover trends and associations for winning chance

Serena

- Perform a very simple analysis to see if there exists any trend with saddlecloth and finish position or price earned
- SexRestriction -> Analysis on C&G group competing in other competition
- StartingLine -> might be associated with handicap distance further study needed
- NoFrontCover -> needs further analysis especially for value -9
- PositionInRunning, wide rail -> needs further analysis (at what moment was the position recorded, scales of Wideoffrail?)
- **Note: all analysis were completed using Rstudio (refer to Rstudio file that has been shared in Github)**