Summary of 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
"Hereos" 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**

Volt start (let the sulkys 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 the first place w/o galloping (galloping is

How are racing organized?

**important: drier's tactics during the race VARY

Here's a simplified summary of the guidebook for the Modelathon contest related to predicting trot racing results in France:

Trot Racing in France:
- Trot racing involves horses racing at a specific trotting gait, either pulling a cart or ridden by a jockey.
- France hosts over 11,000 trot racing events each year, making it a significant hub for trot racing.
- There are two types of trot races in France: harness racing (attele) and mounted racing (monte). The same horse can participate in both types.
- The key roles in preparing a horse for racing are the owner, trainer, and rider. In some cases, one person may perform multiple roles.

Characteristics of French Trot Racing:

- Horses must maintain a trotting gait throughout the race; galloping is not allowed.
- Horses can race from the age of 2 and continue until they are 10 years old.
- France has 56 racetracks for trot racing, with Paris-Vincennes being the most prestigious.
- Racetracks vary in length, and trot races can be held at different distances.

Track Surfaces:
- There are three types of track surfaces: grass (turf), sand, and ash/cinder. Horses may have preferences for different surfaces.

The Start:
- Trot races start in two ways: auto start, where horses line up behind a motorized gate, or volte start, where horses gather on the track and start in an orderly fashion.
- Three factors that affect the start of the race which are
    1. Width of the racecourse
    2. Number of starters
    3. Age category

The Race:
- The goal of the rider is to finish the race as fast as possible while the horse maintains a trot. Galloping leads to disqualification.
- Riders use various tactics to win, including taking the lead early or making a move in the final straight.

Class Restrictions:
- Races have specific conditions based on factors like horse age, earnings, and distance.

Disqualification Rules:
- Horses can be disqualified for breaking stride or other faults during the race.

Weight Restrictions for Riders:
- The rider's weight is important, especially in monte races. There are weight restrictions based on experience.

Horse Shoes:
- Horses can race with horse shoes, plates, or barefoot. Rules govern the use of these options.

Prize Money:

- Prize money is awarded to the top seven finishers in trot races, and the amount reflects the race's quality.

The Data:

- The dataset provided for the contest includes various parameters, both standard and performance-related.
- Standard columns provide information available before a race starts, while performance columns are only available after the race.
- A variable called "winprobability" needs to be created for each observation and should sum to one by RaceID.

Model Training:

- Contestants should exclude data since November 1, 2021, for model training and use October 31 as the last day for training.
    1. Key conditions for winning

Start Position

Racing Unshod (no shoes) -> 10% lower velocity than wearing shoes -> but somehow gives positive impact for performance (preferred over wearing shoes in large tournaments)

Racing unshod increases 25% for chance of being disqualified

Translate the features to everyday language

*Non-significant Predictors are ~~crossed out~~

Colour Codes
- Orange highlight: VERY important
- Italicized : reconsider
- Yellow: figure out later

— 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 (I think…)~~
~~Due to MISSING VALUES~~

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~~

~~"GoingAbbrev" typically refers to an abbreviation or shorthand code used in horse racing to describe the condition of the racecourse, specifically the "going" or the surface underfoot. The going refers to the condition of the racetrack in terms of its firmness or softness, and it can greatly impact a horse's performance. The GoingAbbrev code is a way to quickly convey this information on race programs, forms, or racecards.~~

~~GoingAbbrev codes can vary from one racing authority or region to another, but they generally use a set of standardized abbreviations to describe the condition of the racetrack. Common GoingAbbrev codes include:~~

- ~~GF (Good to Firm): The track is firm and fast, providing good conditions for fast racing.~~
- ~~G (Good): The track is in good condition, but it may not be as firm as "Good to Firm."~~
- ~~GS (Good to Soft): The track is somewhat soft and may slow down the pace of racing.~~
- ~~S (Soft): The track is soft and can be slower, with horses sinking into the ground a bit.~~
- ~~Y (Yielding): The track is soft and wet, and it can be more tiring for the horses.~~
- ~~H (Heavy): The track is very wet and soft, often due to heavy rainfall, making it the slowest and most demanding going.~~

~~GoingID~~

~~=> use either goingID or going abbrev~~

— 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
- *The 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 is 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.

In French horse trot racing, the categorization of horses into different race groups, such as G1 and G2, is based on their performance and achievements. Generally, a horse that belongs to a higher race group, such as G1, is eligible to compete in races of lower groups, including G2 or lower. However, the decision to enter a horse in a particular race group ultimately depends on the owner, trainer, and race organizers, who may consider factors such as the horse's form, ability, and the competitiveness of the race.

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 (most likely it won't affect but we can consider if this affects 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**: Colts are young **male** horses, typically under the age of 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.

When a race or event is designated for "C&G," it means that it is exclusively 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 and competitive racing conditions for male horses, while fillies and mares may have their own separate races.

Quick qs: is it C AND G or C or G?

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

In the context of horse racing, the term "cover" typically refers to a horse's position or placement within the pack 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.

Here's how "cover" is often used in horse racing:

> Position in the Pack: "Cover" may describe a horse's position in relation to other horses during a race. **When a horse is said to have "good cover," it means that it is racing behind or alongside other horses, which can provide it with a favorable position for drafting (following the horses in front) and conserving 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.
> Tactics and Strategy: Jockeys and trainers often strategize to give their horses good cover during a race. Finding a good spot behind other horses can help a horse conserve energy and make a strong move when it matters most, such as in the stretch run.
> Drafting: Drafting behind other horses (using them to block wind resistance) can help a horse maintain its speed and stamina. Good cover is particularly valuable in longer races where conserving energy is essential.

In summary, "cover" in horse racing refers to the positioning of a horse during a race and can be important in terms of race tactics, energy conservation, and the assessment of a horse's performance.

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

Find out what factors are correlated with each other

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 -> maybe 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 wideofrail?)