STAT 847: Analysis Assignment 1

DUE: Friday, February 2 2024 by 11:59pm EST

NOTES

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically in .pdf format via Crowdmark.

Organization and comprehensibility is part of a full solution. Consequently, points will be deducted for solutions that are not organized and incomprehensible. Furthermore, if you submit your assignment to Crowdmark, but you do so incorrectly in any way (e.g., you upload your Question 2 solution in the Question 1 box), you will receive a 5% deduction (i.e., 5% of the assignment's point total will be deducted from your point total).

There are a total of 65 points possible.

For this assignment, you have the data for the 2023 horse racing seasons at Woodbine Racetrack (Toronto), Assiniboia Downs (Winnipeg), and Hastings Park (Vancouver), as well as some of the 2022 season at Woodbine. The “scraped data” dataset is the most detailed of the datasets, where as “afterQ1” is a race-by-race summary.

If you don’t see “afterQ1” in Learn yet, I still have to add it.

The variables are as follows.

Variable Name	Description
meet_location	Racetrack of the event (race)
meet_wday	Weekday of the race
meet_mday	Calendar day of the race
meet_year	Calendar year of the race
racecount	Number of horses in the race
race_number	Race of the day (1st, 2nd, . . .)
horse_number	Number the horse wore during the event
horse_name	Name of the horse
horse_sire	Name of the father of the horse
horse_trainer	Trainer of the horse
horse_jockey	Jockey riding the horse
horse_odds	Stated odds on the horse to win
horse_odds decimal	horse_odds converted to a decimal
horse_place	Place horse finished, 5th or higher is 5, did not run is NA
purse	Total prize money awarded to the owners, jockeys, and trainers
time_frac1	Time when the lead horse finishes the 1st fraction of the race
time_frac2	Time when the lead horse finishes the 2nd fraction of the race
time_frac3	Time when the lead horse finishes the 3rd fraction of the race
time_frac4	Time when the lead horse finishes the 4th fraction of the race
time_frac5	Time when the lead horse finishes the 5th fraction of the race
time_final	Time when the lead horse finishes the race, in seconds
track_length	Distance of the entire race in (F)urlongs or (M)iles
track_type	Type of ground. (Dirt, Turf, All Weather Track)
race_class	Additional information about the race
dist_frac1	Distance from start to 1st fraction
dist_frac2	Distance from start to 2nd fraction
dist_frac3	Distance from start to 3rd fraction
dist_frac4	Distance from start to 4th fraction
dist_frac5	Distance from start to 5th fraction

There is also a code file, “HRN EDA.txt” as well as a presentation based on this data titled “How to start a sports analytics project.html” that may give additional context, as well as code you can use for this assignment.

Q1. (10 marks) Currently in each of three “scraped data” datasets, one row represents one horse. Make a new dataset for the Woodbine dataset where one line represents one race instead. Keep all the variables pertaining to the race and drop the ones pertaining to the horse (that is, drop `horse_number`,`horse_name`,`horse_sire`,`horse_trainer`,`horse_jockey`,`horse_odds`,`horse_odds_decimal`, and `horse_place`.)

Use the mini case study that uses one large `ddply()` function as a basis for your code.

Show your code and the first 3 rows of the new dataset.

Since this question makes other, later questions easier, you may use the “afterQ1” datasets for Q2 onward.

Q2. (5 marks) At Woodbine, calculate the average time it takes for the winning horse to complete a race of each available length (hint: use the by() command). Present your answer as a table like the following, and round average times to two decimal places.

Event length	Average Time
3F (3 Furlongs)	53.25
6F	101.42
1M (1 Mile, 8 Furlongs)	

Q3. (5 marks) At Woodbine, find the probability of a horse coming in second place as a function of the decimal odds, rounded to the nearest whole number. Present your answer as a table like the following, and round probabilties to three decimal places. You may use the provided EDA code for the first place probabilities as a starting point.

Rounded Odds	Probability of 2nd
1	0.142
2	0.241
3	

Q4. (6 marks) At Woodbine, conduct a two-sample t-test to see if the finish times differ on average between turf tracks and the all weather track for 6F (6 furlong) length races. For this question, assume that 'inner turf' and 'turf' are both turf tracks that belong in the same group. Use alpha = 0.05 as your cut-off for significance.

Q5. (8 marks) Make a side-by-side boxplot of the finish times for 6F races between the three locations. That is, make a boxplot where each of the three boxes shows the distribution of times from Woodbine, Assiniboia, or Hastings. Either base R or ggplot is acceptable.

Q6. (8 marks) Find the names of the five horses that have won the most evnets at Woodbine (in 2022 and 2023 combined) and their total number of wins. Present their results in a table like so.

Horse	Wins
Rainbow Dash	11
Twilight Sparkle	7
The cowboy one	6

Q7. (8 marks) Typically, a purse is divided so that 60% goes to the winner, 20% goes to 2nd place, 10% goes to 3rd place, and the remaining 10% is split among all the other horses that finish. Assume that this purse payout system is used at Woodbine. Find the names of the five horses that have won the most money at Woodbine (in 2022 and 2023 combined) and their total winnings during these two years.

Horse	Prize Money
Rainbow Dash	654,000
Twilight Sparkle	321,000
The cowboy one	

Q8. (10 marks) Every race has fractional times, which are the times when the leading horse finishes some fraction of the race. For every race that is between 4 1/2F and 1 9/16M inclusive, the second fraction (`time_frac2`) is the time that the first horse finishes 1/2 a mile (4 furlongs).

Plot as a broken line plot of `time_frac2` as a function of distance for all the distances between 4 1/2F and 1 9/16M. Be sure to convert the distances into something numeric like number of furlongs; 1 mile is 8 furlongs.

Q9. (5 marks) Fit a quadratic model using `lm()` of `time_frac2` as a function of distance for all the distances between 4 1/2F and 1 9/16M. Be sure to convert the distances into something numeric like number of furlongs; 1 mile is 8 furlongs. Report the `summary()` of the model.