# Horse Racing and Exploratory Data Analysis

Jack Davis

In this lesson, we are given a large dataset, in this case of horse race results, and our job is to use R to extract some basic facts about the dataset.

## The PPDAC Model

We can employ the PPDA model.

- Problem: Describe what you want to do.
- Plan: Figure out what you're doing while mistakes are cheap.
- Data: Actually gather the data
- Analysis: Turn data into insights.

## Problem - What to do?

Problem (General) - We want to diversify my sports analytics base, horseplaying is a blind spot.

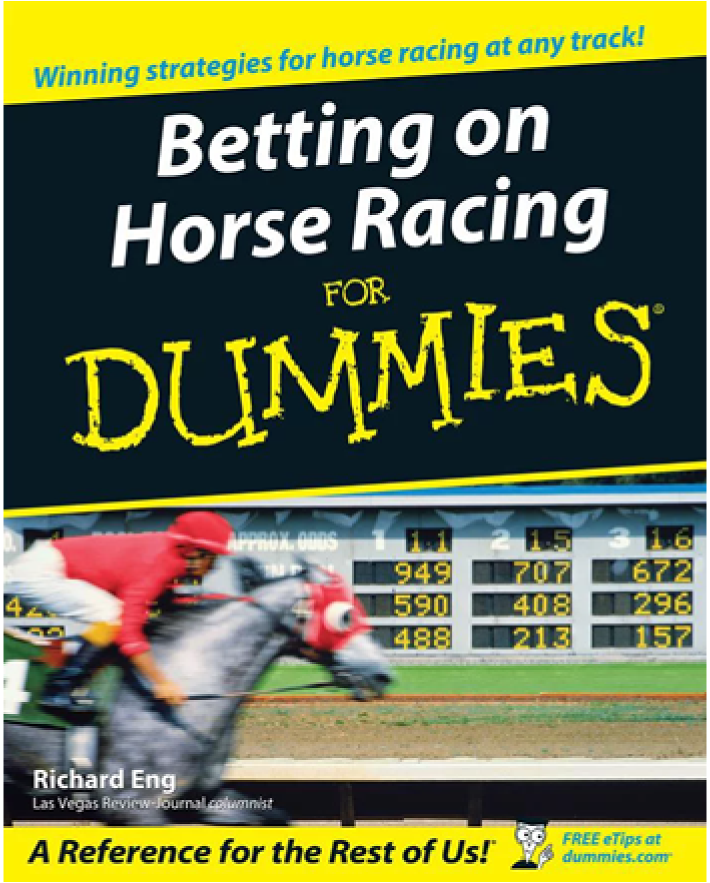
Problem (Specific) - We want to estimate the probability that different horses win races.

Problem (More specific) - estimate racing probabilities at Woodbine Racetrack in Toronto.

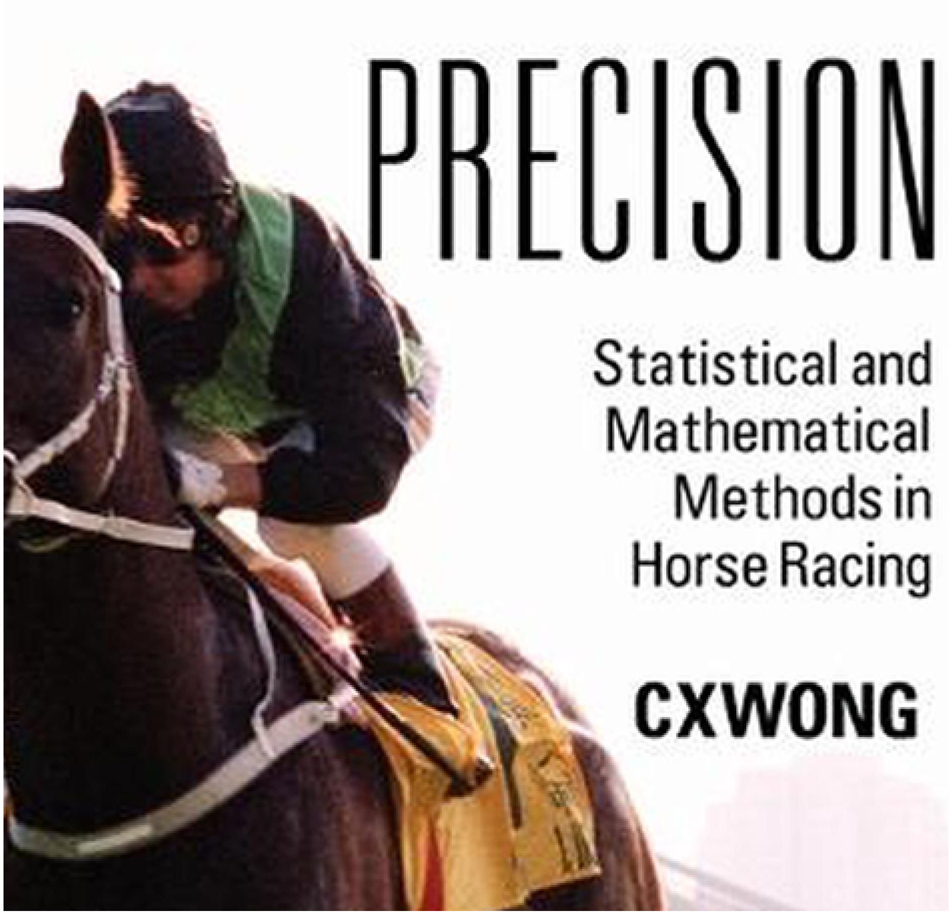
## Plan - Subject knowledge

Plan - What subject knowledge do We need first?

- What the popular models that people try? Can we recreate them to tinker with them or improve?
- Are there any interesting unanswered questions in the field?
- What are the different betting options?
- What is a perimutuel market?
- What is lazix?
- Is track effect worth looking into?
- Is jockey effect worth looking into?


Winning strategies for horse racing at any track!
Betting on Horse Racing
FOR
DUMMIES
APPROX. ODDS
1 1-1
2 1-5
3 1-6
949
707
672
590
408
296
488
213
157
Richard Eng
Las Vegas Review-Journal columnist
A Reference for the Rest of Us!
FREE eTips at dummies.com

The “For Dummies” series is great for getting the basics of many different topic. This book contains a lot of information about the horse betting industry in the United States. It also explains how perimutuel betting works, and the basic variables to look for like jockey effects, different track conditions, and lazix usage.



"Precision" contains information on popular modelling methods for horse racing, the statistical and programming background necessary to use the models, and some information on the horse racing industry in Hong Kong.

Plan - What subject knowledge do We need first?

- What the popular models that people try? Can we recreate them to tinker with them or improve? (Surprisingly ad-hoc)
- Are there any interesting unanswered questions in the field? (Lots of exotic bets are suboptimal)
- What are the different betting options? (Win, Place, Show, Trifecta, Pick-3)
- What is a perimutuel market? (A bettor-vs-bettor market of odds, rather than bettor-vs-house)
- What is lazix? (A drug to reduce internal bleeding, sometimes prescribed to horses, performance enhancing)
- Is track effect worth looking into? (Yes, but how is complex)
- Is jockey effect worth looking into? (Yes, and how is linear)

## Data - Collection Premade

- We need to see what data is available before we build a model. There's no use in relying on variables we can't get.

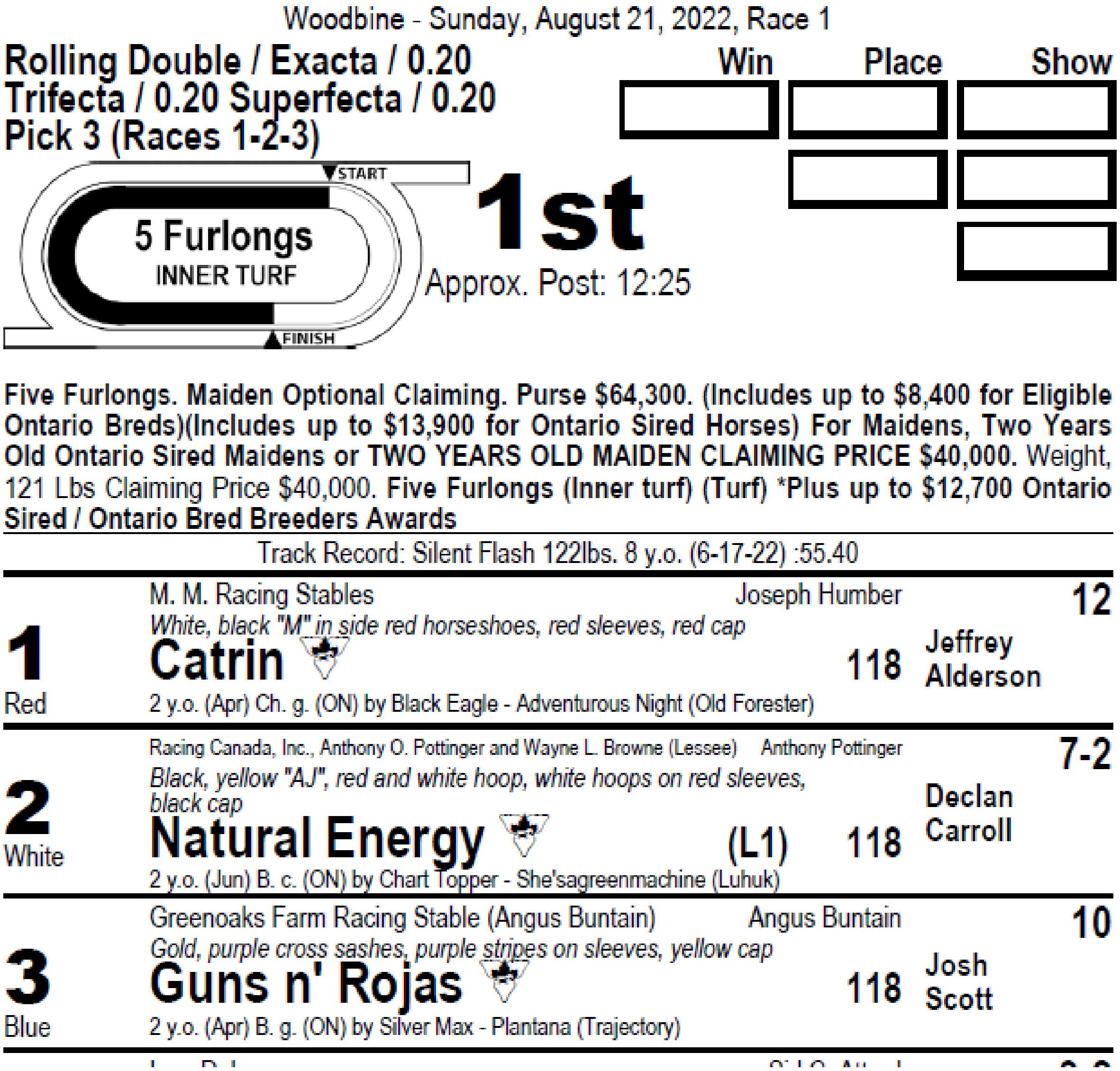
- Someone else's pre-cleaned, pre-formatted data. Great for replicability. Not so great for getting a personal edge.

- https://horseracingdatasets.com is a great start, but it doesn't have Woodbine race results

- Daily Racing Form https://www.drf.com/ has excellent race results and data, but it costs $100 USD/month for 'unlimited' access (and a sports analytics version of 'unlimited' might violate some terms of service)
- There are no 'horse racing' packages on CRAN

## Data - Collection from racing programs

- We want something that's available for many different racetracks, even though I'm starting this project on only one racetrack (START SMALL, LIKE A SINGLE RACE IF YOU HAVE TO)
- Woodbine has programs that describe the races and the horses/riders/trainers/owners in detail. It's in PDF so we'd have to OCR (Optical Character Recognition, with the `tesseract` package in R) it.
- Many racetracks have very similar programs on PDFs, so having a system to read them would be great and possibly scalable.



Woodbine - Sunday, August 21, 2022, Race 1

**Rolling Double / Exacta / 0.20**
**Trifecta / 0.20 Superfecta / 0.20**
**Pick 3 (Races 1-2-3)**

**Win** **Place** **Show**

START
5 Furlongs
INNER TURF
FINISH

**1st**

Approx. Post: 12:25

**Five Furlongs. Maiden Optional Claiming. Purse $64,300. (Includes up to $8,400 for Eligible Ontario Breds)(Includes up to $13,900 for Ontario Sired Horses) For Maidens, Two Years Old Ontario Sired Maidens or TWO YEARS OLD MAIDEN CLAIMING PRICE $40,000.** Weight, 121 Lbs Claiming Price $40,000. **Five Furlongs (Inner turf) (Turf) *Plus up to $12,700 Ontario Sired / Ontario Bred Breeders Awards**

Track Record: Silent Flash 122lbs. 8 y.o. (6-17-22) :55.40

**1** Red
M. M. Racing Stables — Joseph Humber — **12**
*White, black "M" in side red horseshoes, red sleeves, red cap*
**Catrin** — 118 — **Jeffrey Alderson**
2 y.o. (Apr) Ch. g. (ON) by Black Eagle - Adventurous Night (Old Forester)

**2** White
Racing Canada, Inc., Anthony O. Pottinger and Wayne L. Browne (Lessee) — Anthony Pottinger — **7-2**
*Black, yellow "AJ", red and white hoop, white hoops on red sleeves, black cap*
**Natural Energy** — (L1) — 118 — **Declan Carroll**
2 y.o. (Jun) B. c. (ON) by Chart Topper - She'sagreenmachine (Luhuk)

**3** Blue
Greenoaks Farm Racing Stable (Angus Buntain) — Angus Buntain — **10**
*Gold, purple cross sashes, purple stripes on sleeves, yellow cap*
**Guns n' Rojas** — 118 — **Josh Scott**
2 y.o. (Apr) B. g. (ON) by Silver Max - Plantana (Trajectory)

OCR works best when words are typed clearly (which they mostly are here), and in neat lines of the same size (which they are not)

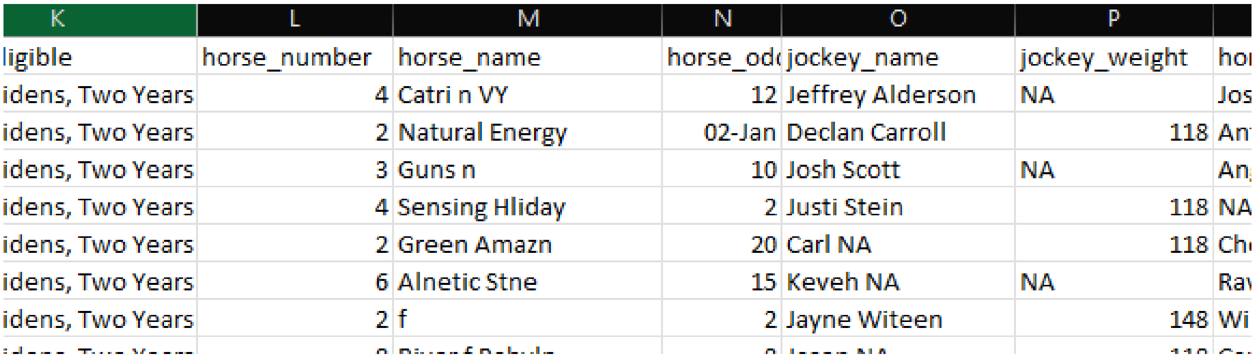
Text features can be extracted with regular expression-based functions from the `stringr` package.

```
race_wday = str_extract(raw_race[1], "[a-zA-Z]+day")
race_mday = str_extract(raw_race[1], "[a-zA-Z]+ [0-9]{1,2}")
race_year = str_extract(raw_race[1], "20[0-9][0-9]")
race_number = str_extract(raw_race[1], "Race [0-9]{1,2}")
race_number = str_replace(race_number, "^Race ", "")
```

We can also clean OCR data with functions from the same package

```
## Remove fancy apostrophe
raw_all = str_replace_all(raw_all, "’", "'")
raw_all = str_replace_all(raw_all, """, "\"")
raw_all = str_replace_all(raw_all, "[©=]", "")
```

```
JWidx = which(str_detect(raw_horse, "1[0-9 ]+"))[1]
jockey_weight = str_extract(raw_horse[JWidx], "1[0-9 ][0-9]")
jockey_weight = str_replace_all(jockey_weight, " ", "")
```

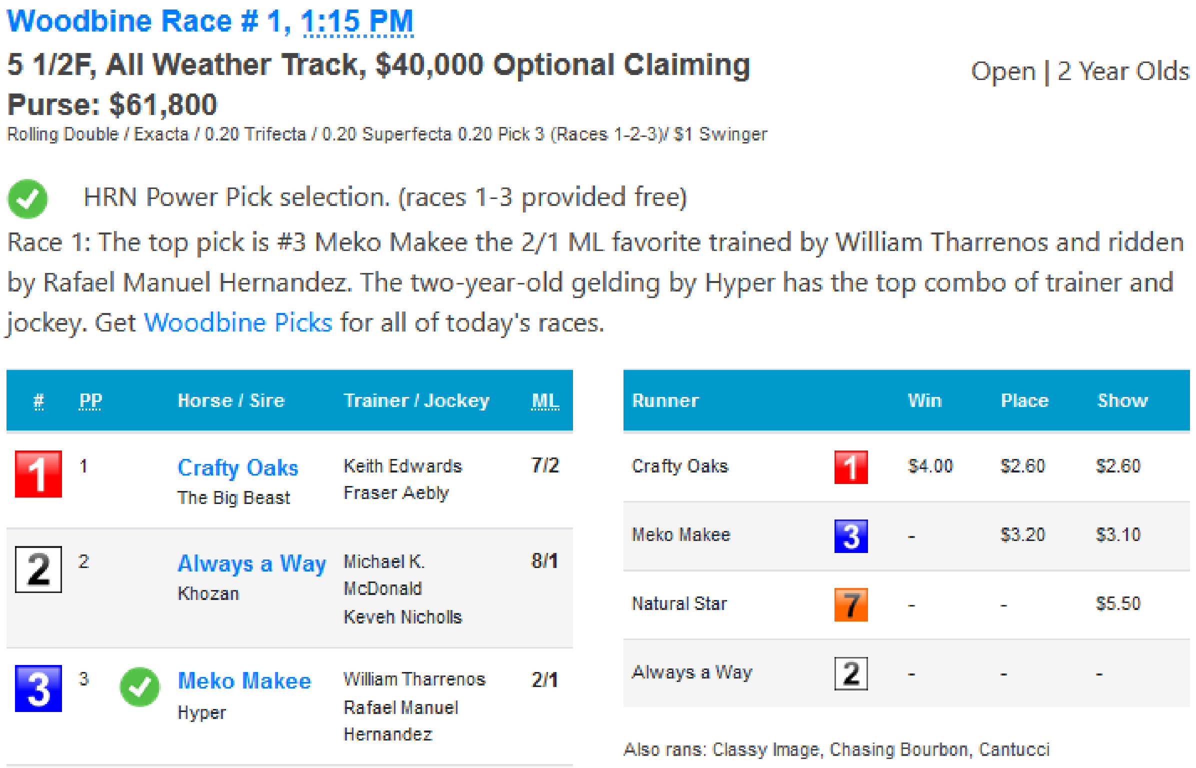


| K | L | M | N | O | P | |
|---|---|---|---|---|---|---|
| ligible | horse_number | horse_name | horse_od | jockey_name | jockey_weight | ho |
| idens, Two Years | 4 | Catri n VY | 12 | Jeffrey Alderson | NA | Jos |
| idens, Two Years | 2 | Natural Energy | 02-Jan | Declan Carroll | 118 | An |
| idens, Two Years | 3 | Guns n | 10 | Josh Scott | NA | An |
| idens, Two Years | 4 | Sensing Hliday | 2 | Justi Stein | 118 | NA |
| idens, Two Years | 2 | Green Amazn | 20 | Carl NA | 118 | Ch |
| idens, Two Years | 6 | Alnetic Stne | 15 | Keveh NA | NA | Ra |
| idens, Two Years | 2 | f | 2 | Jayne Witeen | 148 | Wi |

Even then, it was too inconsistent to use. According to the ‘fail faster, fail cheaper’ strategy, time for a new approach.

## Data - Horse Racing Nation

- https://www.horseracingnation.com



Woodbine Race # 1, 1:15 PM

5 1/2F, All Weather Track, $40,000 Optional Claiming

Purse: $61,800

Open | 2 Year Olds

Rolling Double / Exacta / 0.20 Trifecta / 0.20 Superfecta 0.20 Pick 3 (Races 1-2-3)/ $1 Swinger

✓ HRN Power Pick selection. (races 1-3 provided free)

Race 1: The top pick is #3 Meko Makee the 2/1 ML favorite trained by William Tharrenos and ridden by Rafael Manuel Hernandez. The two-year-old gelding by Hyper has the top combo of trainer and jockey. Get Woodbine Picks for all of today's races.

| # | PP | | Horse / Sire | Trainer / Jockey | ML |
|---|---|---|---|---|---|
| 1 | 1 | | Crafty Oaks<br>The Big Beast | Keith Edwards<br>Fraser Aebly | 7/2 |
| 2 | 2 | | Always a Way<br>Khozan | Michael K. McDonald<br>Keveh Nicholls | 8/1 |
| 3 | 3 | ✓ | Meko Makee<br>Hyper | William Tharrenos<br>Rafael Manuel Hernandez | 2/1 |

| Runner | | Win | Place | Show |
|---|---|---|---|---|
| Crafty Oaks | 1 | $4.00 | $2.60 | $2.60 |
| Meko Makee | 3 | - | $3.20 | $3.10 |
| Natural Star | 7 | - | - | $5.50 |
| Always a Way | 2 | - | - | - |

Also rans: Classy Image, Chasing Bourbon, Cantucci

Horse racing nation has results in tables on the web.

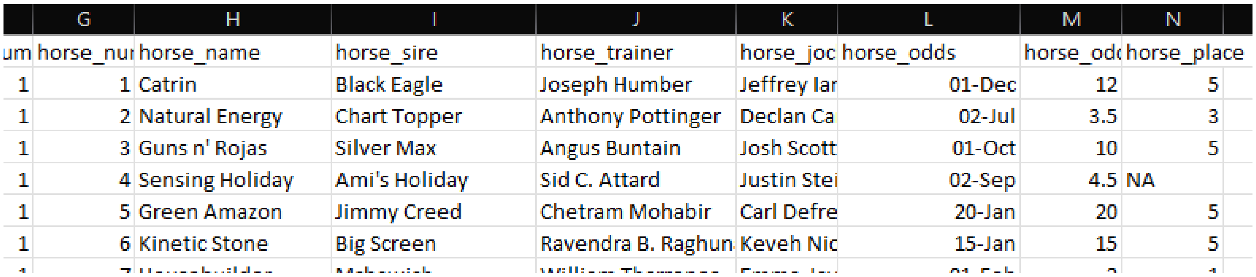
We can try to scrape those with `rvest`, but their website has anti-bot tech.

So we'll do it directly with a mouse-and-keyboard macro using Asoftech Automation:

- 1. Take first URL from a list in a notepad,
- 2. copy into address bar,
- 3. wait for page to load,
- 4. wait some more because page load times have variance,
- 5. crtl + A, then crtl + C to grab all the text
- 6. crtl + V in a notepad

Record steps 1-6, set to repeat 200 times.

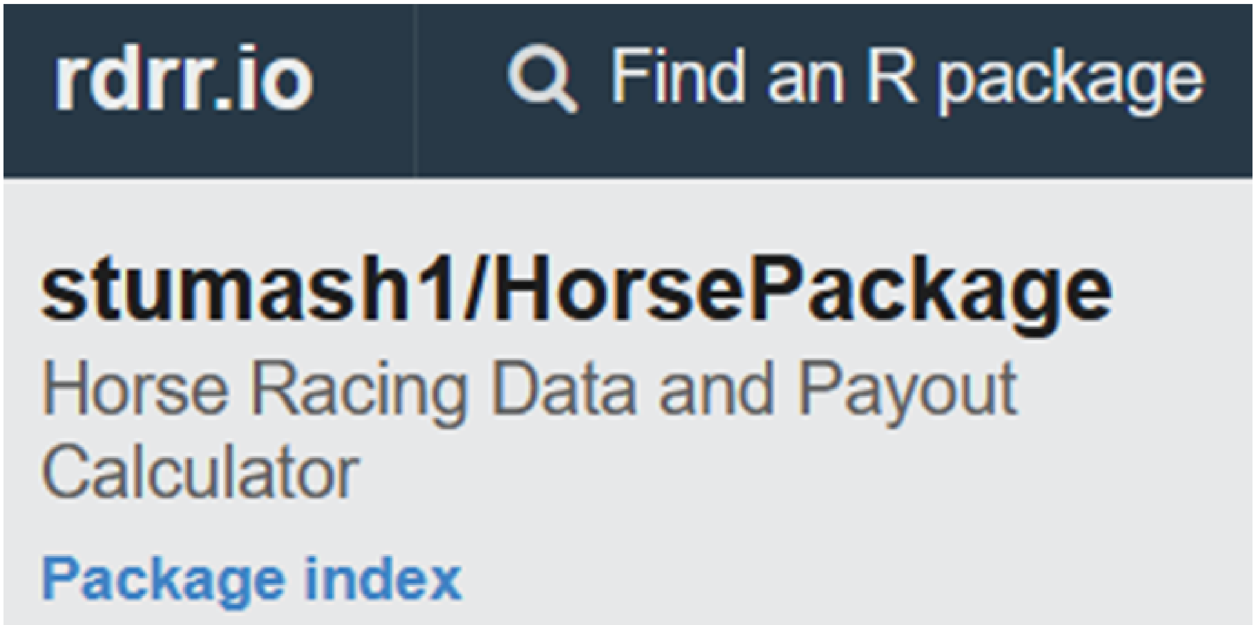
Using similar text extraction and cleaning functions.



| | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|
| um | horse_nu | horse_name | horse_sire | horse_trainer | horse_joc | horse_odds | horse_od | horse_place |
| 1 | 1 | Catrin | Black Eagle | Joseph Humber | Jeffrey Iar | 01-Dec | 12 | 5 |
| 1 | 2 | Natural Energy | Chart Topper | Anthony Pottinger | Declan Ca | 02-Jul | 3.5 | 3 |
| 1 | 3 | Guns n' Rojas | Silver Max | Angus Buntain | Josh Scott | 01-Oct | 10 | 5 |
| 1 | 4 | Sensing Holiday | Ami's Holiday | Sid C. Attard | Justin Stei | 02-Sep | 4.5 | NA |
| 1 | 5 | Green Amazon | Jimmy Creed | Chetram Mohabir | Carl Defre | 20-Jan | 20 | 5 |
| 1 | 6 | Kinetic Stone | Big Screen | Ravendra B. Raghun | Keveh Nic | 15-Jan | 15 | 5 |
| 1 | 7 | Housebuilder | Mahowish | William Tharrenos | Emma-Jay | 01-Feb | 2 | 1 |

We don't have all the details of the program, but we can build upon this later with a data merge if we want.

## Data - Horsepackage



stumash1/HorsePackage

Horse Racing Data and Payout Calculator

Stumash's Horsepackage calculates box odds and trifecta probabilities using the Harville method, and has some sample data as well. Handy for testing things, and for future analyses. (Again, try to avoid redoing others' work if you can.)

https://rdrr.io/github/stumash1/HorsePackage/

## Anaysis - Exploratory Data Analysis

What can we learn very quickly from the Horse Racing Nation dataset?

First look at the data

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 4.3.2
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
hrn = read.csv("HRN scraped data 2023-10-11.csv")
head(hrn)
```

```
##   meet_location meet_wday meet_mday meet_year racecount race_number
## 1      Woodbine    Sunday August 21      2022         1           1
## 2      Woodbine    Sunday August 21      2022         1           1
## 3      Woodbine    Sunday August 21      2022         1           1
## 4      Woodbine    Sunday August 21      2022         1           1
## 5      Woodbine    Sunday August 21      2022         1           1
## 6      Woodbine    Sunday August 21      2022         1           1
##   horse_number      horse_name    horse_sire         horse_trainer
## 1            1          Catrin   Black Eagle         Joseph Humber
## 2            2  Natural Energy  Chart Topper     Anthony Pottinger
## 3            3   Guns n’ Rojas    Silver Max         Angus Buntain
## 4            4 Sensing Holiday Ami’s Holiday         Sid C. Attard
## 5            5    Green Amazon   Jimmy Creed       Chetram Mohabir
## 6            6   Kinetic Stone    Big Screen Ravendra B. Raghunath
##           horse_jockey horse_odds horse_odds_decimal horse_place
## 1 Jeffrey Ian Alderson       12/1               12.0           5
## 2       Declan Carroll        7/2                3.5           3
## 3           Josh Scott       10/1               10.0           5
## 4         Justin Stein        9/2                4.5          NA
## 5       Carl Defreitas       20/1               20.0           5
## 6       Keveh Nicholls       15/1               15.0           5
```

Next, how do the winning odds change with final place.

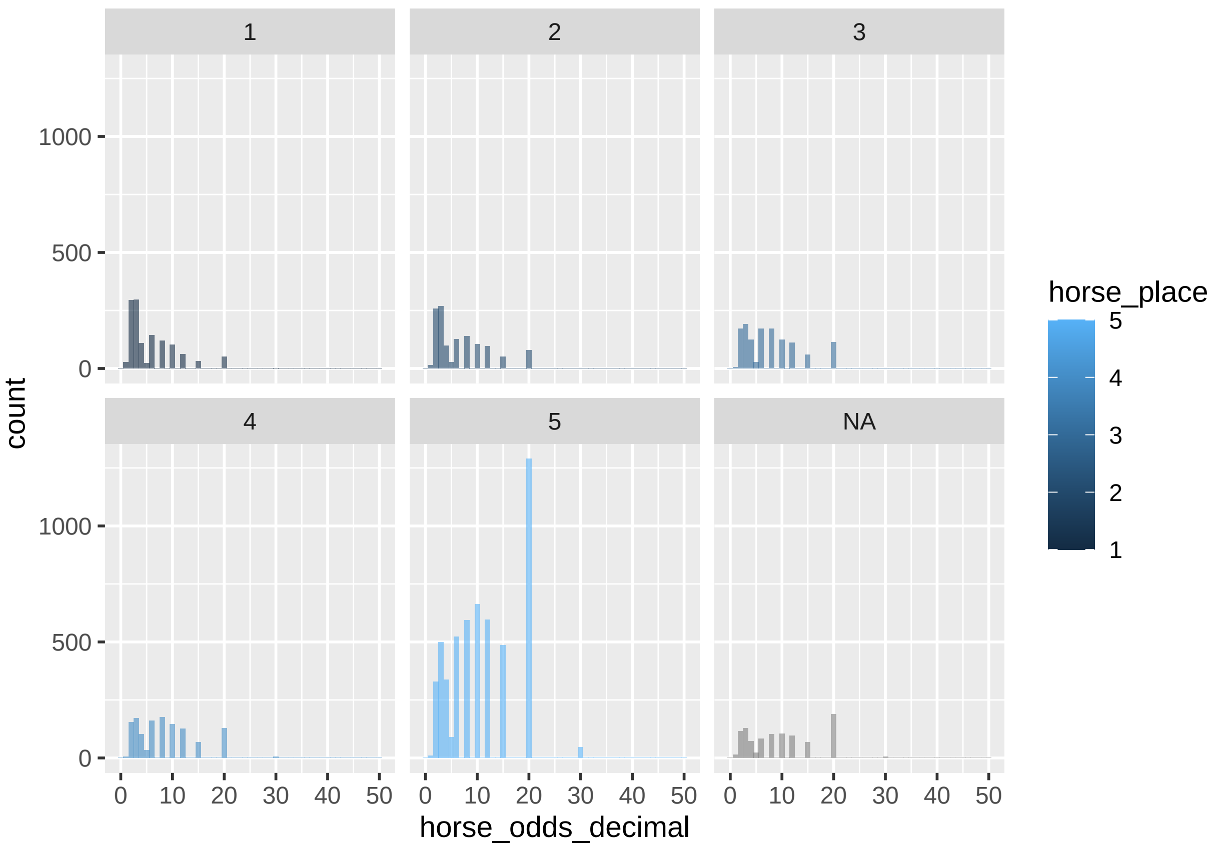
```
ddply(hrn, "horse_place", summarize,
      n = length(horse_odds_decimal),
      median_odds = median(horse_odds_decimal),
      mean_odds = mean(horse_odds_decimal),
      sd_odds = sd(horse_odds_decimal))
```

```
##   horse_place    n median_odds mean_odds  sd_odds
## 1           1 1259         4.0  5.755997 4.453432
## 2           2 1260         4.5  6.630556 4.968725
## 3           3 1267         6.0  7.706709 5.228626
## 4           4 1274         8.0  8.284458 5.491038
## 5           5 5466        10.0 11.054446 6.449560
## 6          NA 1004         8.0  9.648506 6.720582
```

They do change. Will a set of histograms help?

```
p1 <- ggplot(hrn, aes(x=horse_odds_decimal, fill=horse_place)) +
    geom_histogram(alpha=0.6, binwidth = 1) +
    facet_wrap(~horse_place)

p1
```



Let's flip the problem on its head. Looking at distribution of place as a function of decimal odds

```
hrn$floor_odds = floor(hrn$horse_odds_decimal)

tab1 = table(hrn$floor_odds, hrn$horse_place)
tab1
```

```
##
##         1    2    3    4    5
##   0     4    1    1    0    1
##   1    75   44   25   18   41
##   2   244  228  150  141  297
##   3   297  268  191  172  500
##   4   108   98  123  103  338
##   5    23   27   27   33   90
##   6   143  127  172  160  523
##   8   119  139  171  175  594
##   10  102  105  124  145  662
##   12   62   95  110  126  596
##   15   31   50   59   68  487
##   20   50   78  114  128 1289
##   30    1    0    0    5   46
##   50    0    0    0    0    2
```

Floor of zero? That means you'd win less than you risked (plus you money back). Is that a mistake?

```
hrn[which(hrn$horse_odds_decimal < 1),]
```

```
##       meet_location meet_wday    meet_mday meet_year racecount race_number
## 2715       Woodbine    Sunday   October 23      2022       303           2
## 4241       Woodbine  Saturday  November 26      2022       465           7
## 5247       Woodbine    Friday        May 5      2023       579           8
## 6367       Woodbine  Saturday       June 3      2023       714           5
## 7248       Woodbine    Sunday      June 25      2023       813           5
## 8350       Woodbine    Sunday      July 30      2023       937          11
## 8631       Woodbine    Friday     August 4      2023       966           6
## 9558       Woodbine    Friday    August 25      2023      1073           6
## 10615      Woodbine    Sunday September 17      2023      1201           3
##       horse_number         horse_name        horse_sire          horse_trainer
## 2715             2 Souper Hoity Toity          Uncle Mo          Mark E. Casse
## 4241             7   War Bomber (IRE)         War Front        Norman McKnight
## 5247             4 Canadiansweetheart   Ransom the Moon         Martin Drexler
## 6367             1              Moira       Ghostzapper           Kevin Attard
## 7248             4            Loyalty         Hard Spun          Josie Carroll
## 8350             5 Patches O’Houlihan            Reload       Robert P. Tiller
## 8631             2             Cotton    Twirling Candy         Martin Drexler
## 9558             2 Reservenotattained    Shanghai Bobby         Martin Drexler
## 10615            5          Oscarsson Oscar Performance Catherine Day Phillips
##                  horse_jockey horse_odds horse_odds_decimal horse_place
## 2715         Patrick Husbands        3/5                0.6           1
## 4241             Sahin Civaci        4/5                0.8          NA
## 5247           Kazushi Kimura        2/5                0.4           1
## 6367           Kazushi Kimura        2/5                0.4           2
## 7248           Kazushi Kimura        4/5                0.8           1
## 8350         Daisuke Fukumoto        4/5                0.8           1
## 8631  Rafael Manuel Hernandez        4/5                0.8           5
## 9558           Kazushi Kimura        4/5                0.8          NA
## 10615          Kazushi Kimura        4/5                0.8           3
##       floor_odds
## 2715           0
## 4241           0
## 5247           0
## 6367           0
## 7248           0
## 8350           0
## 8631           0
## 9558           0
## 10615          0
```

No, Kazushi Kimura is just really REALLY good as a jockey.

Now let’s look at the outcomes as a proportion of the horses

```
tab2 = round(prop.table(tab1, 1),3)
tab2
```

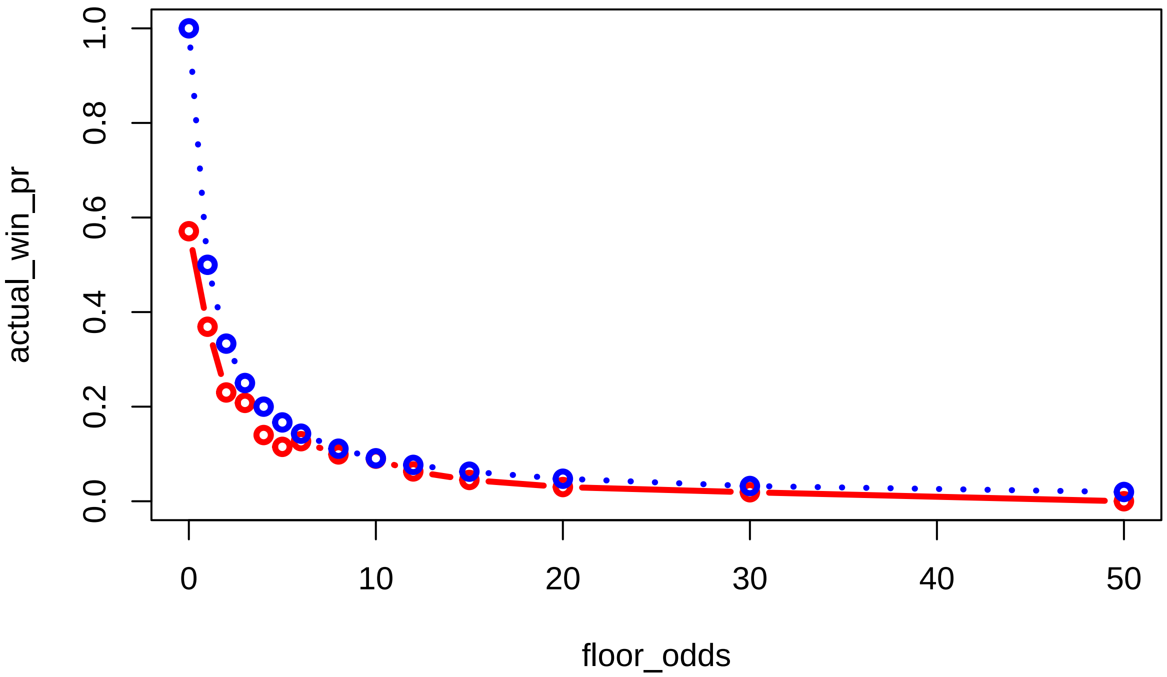
```
##
##          1     2     3     4     5
##   0  0.571 0.143 0.143 0.000 0.143
##   1  0.369 0.217 0.123 0.089 0.202
##   2  0.230 0.215 0.142 0.133 0.280
##   3  0.208 0.188 0.134 0.120 0.350
##   4  0.140 0.127 0.160 0.134 0.439
##   5  0.115 0.135 0.135 0.165 0.450
##   6  0.127 0.113 0.153 0.142 0.465
##   8  0.099 0.116 0.143 0.146 0.496
##   10 0.090 0.092 0.109 0.127 0.582
##   12 0.063 0.096 0.111 0.127 0.603
##   15 0.045 0.072 0.085 0.098 0.701
##   20 0.030 0.047 0.069 0.077 0.777
##   30 0.019 0.000 0.000 0.096 0.885
##   50 0.000 0.000 0.000 0.000 1.000
```

So the horses that pay less than 1/1 (plus your $1 back) win 57.1% of the time. The horses that pay between 1/1 and 2/1 win 36.9% of the time, and so on.

Let's plot this, and overlay 1/(x+1) to it as well because that's the "implied probability" in otherwords, if the odds paid out were "fair, that's what the win probability would be.

```
actual_win_pr = tab2[,1]
implied_win_pr = 1/(as.numeric(row.names(tab2)) + 1)
floor_odds = as.numeric(row.names(tab2))

plot(floor_odds, actual_win_pr, type="b", lwd=3, col="Red", ylim=c(0,1))
lines(floor_odds, implied_win_pr, type="b", lwd=3, col="Blue", lty=3)
```

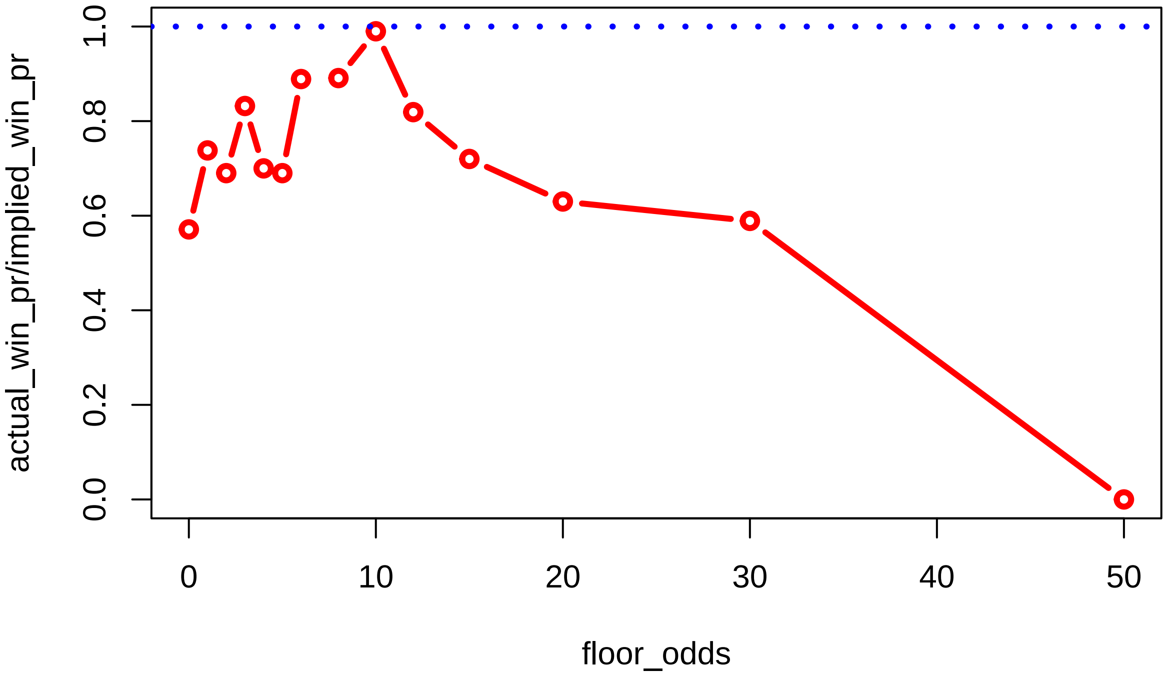


The blue line is ‘fair’ winning prob at a horse racetrack, the red line is the actual winning prob. The fact that the red line is always a bit below the blue line shows the consistent house advantage, which is unsurprizing.

Because we took the floor of each odds offered, not the actual odds, the house advantage is exaggerated, especially at the low odds (the short odds).

We can also get a ratio of actual to implied odds

```
plot(floor_odds, actual_win_pr / implied_win_pr, type="b", lwd=3, col="Red", ylim=c(0,1))
abline(h=1, lty=3, lwd=3, col="Blue")
```



It looks like horses that pay between 7-to-1 and 10-to-1 are the best bets.

However, we don’t know the actual odds until just before race time, so this strategy is hard to refine. If we could predict the odds that would be paid out in advance (which we can model), and fit the probability of each horse winning each race (which we can also model), then we could look for discrepancies where we expect a horse to pay a better than fair price. This graph suggests we would find such horses in the 7-to-1 to 10-to-1 range.

Note that this isn’t the horses that most likely to win, they’re the horses that have the largest payouts relative to their chance to win.