Horse Racing

Abhishek Baral

12/10/2019

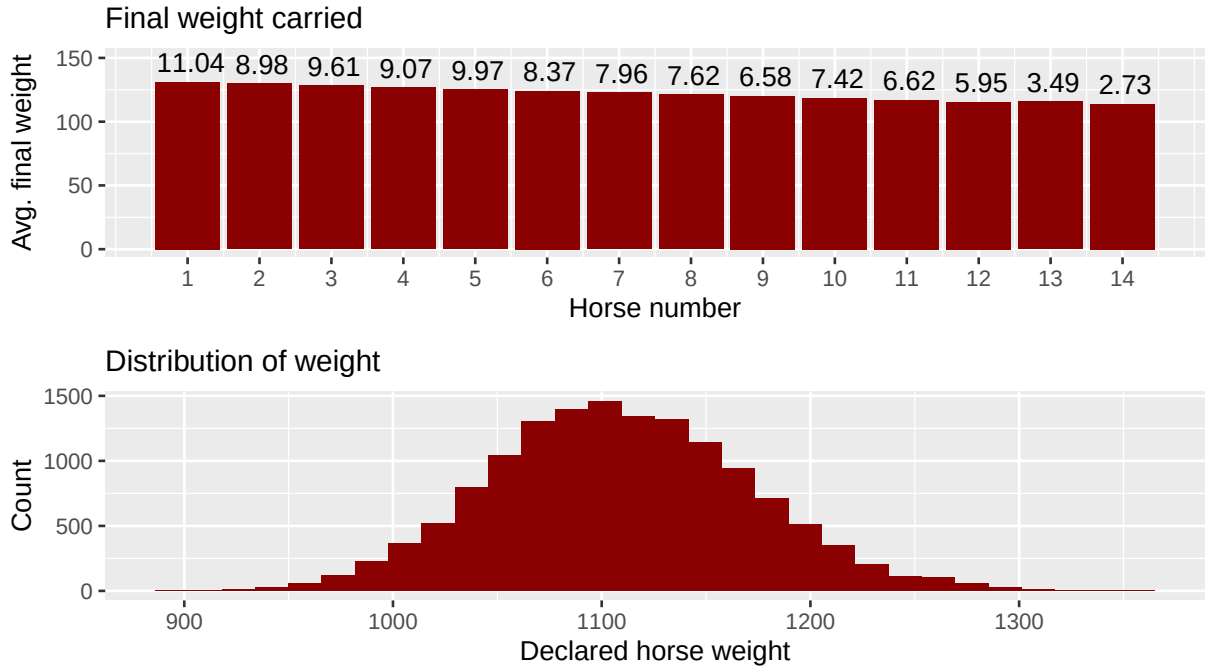
Abstract

This report investigates what factors are relevant in placing in the top 3 positions in a heat and see if we can create a classifier to predict future horse winnings. Using data from the Hong Kong Jockey Club we were able to visualize and understand relationships between the predictors and the target variable, top 3. A logit and random forest model were used to predict future horse winnings through a train and test split. Unfortunately, neither model was able to score high enough in sensitivity for it to be considered valid. Since the dataset spans a little over two years, further testing should be done on more years of data as well as looking into other classifiers.

Intro

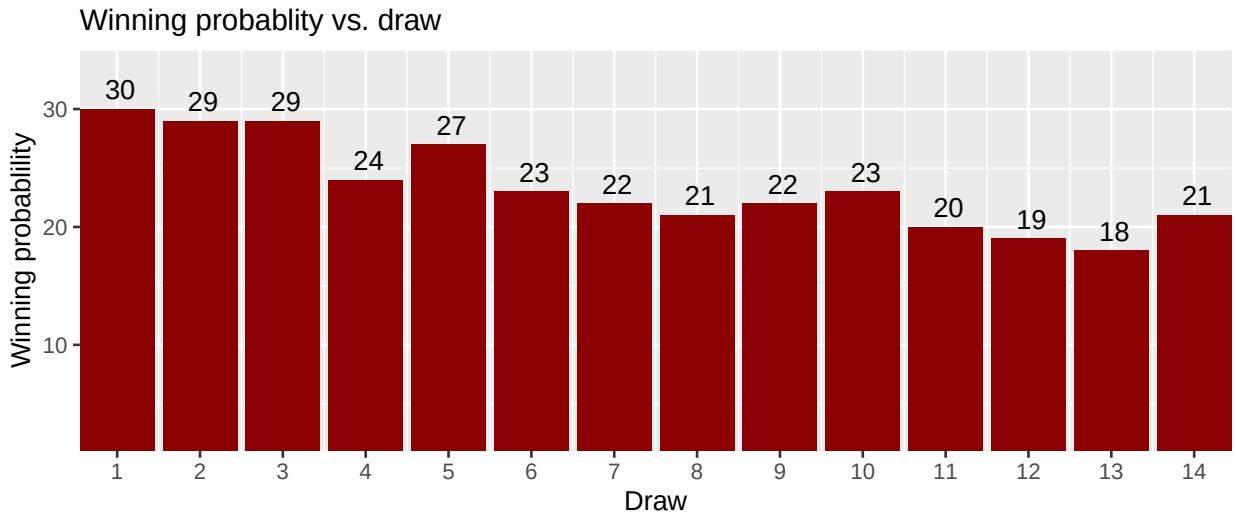
This paper was inspired by Bill Benter, who developed on of the most successful statistical softwares in predicting horse race winnings for the Hong Kong racing market, nearly raking in 1 billion dollars during his career. At the time, Bill's model used 16 variables that could affect the horses, jockeys, and trainers of a race and its impact on the outcome of the race. On average, he brought home between 5 to 10 million dollars in a single day at the racetrack during his heyday. The purpose of this paper is to attempt to recreate an earlier version of Bill's model, in the hope of understanding how machine learning can be used in the area of sports betting. The paper will dive into a few machine learning methods in the hopes that a certain method will be able to predict winning horses better than chance.

EDA/Analysis

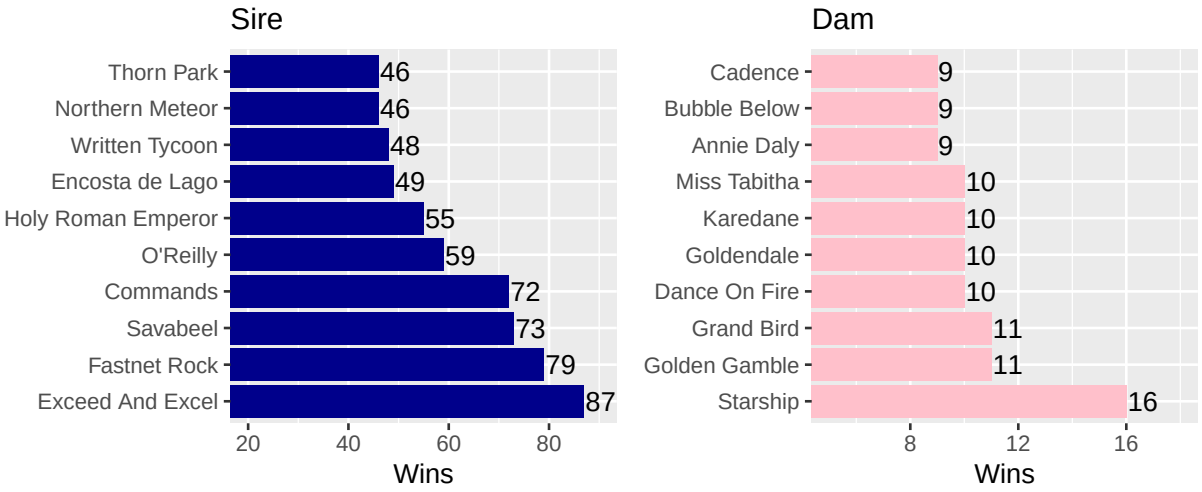


In horse racing a handicap system is used where each horse is allocated a weight based on its ability in an attempt to equalize every horses' chances of winning. A better horse indicated by a lower horse number will carry more weight since the handicapper believes they are more likely to win the race. Looking at the top graph, the handicapper has done somewhat of a decent job in attempting to equalize the odds of winning. On average, the most favored horse to win carries an additional 131 pounds of extra weight and has an 11.04% chance of winning, while the least favored wears an additional 119 pounds with a 2.73% chance of winning. The handicapper should put even more weight on the number 1 horse while taking off weight for the number 14 horse to further equalize the playing field.

Moving to the lower graph, we see that the distribution of the horse weight is surprisingly normal, with a mean of 1108 pounds before the additional weight tacked on and a standard deviation of 62 pounds. This went against my person expectation, as I would have assumed that the distribution would be bimodal, where stronger/heavier horses would be used in shorter length races and lighter horses would be more common for longer length races.



The draw or gate number is drawn two days before the races. Gate number 1 is the inner most lane, so we would expect horses have lower draws to have higher chances of winnings. We see that the first gate has the highest probability of winning with roughly a 30 percent chance on average. What is interesting is that there is there is no linear decrease as draw increases, for example, draw 5 has the fourth highest chance of winning and draw 13 is the worst draw rather than 14.

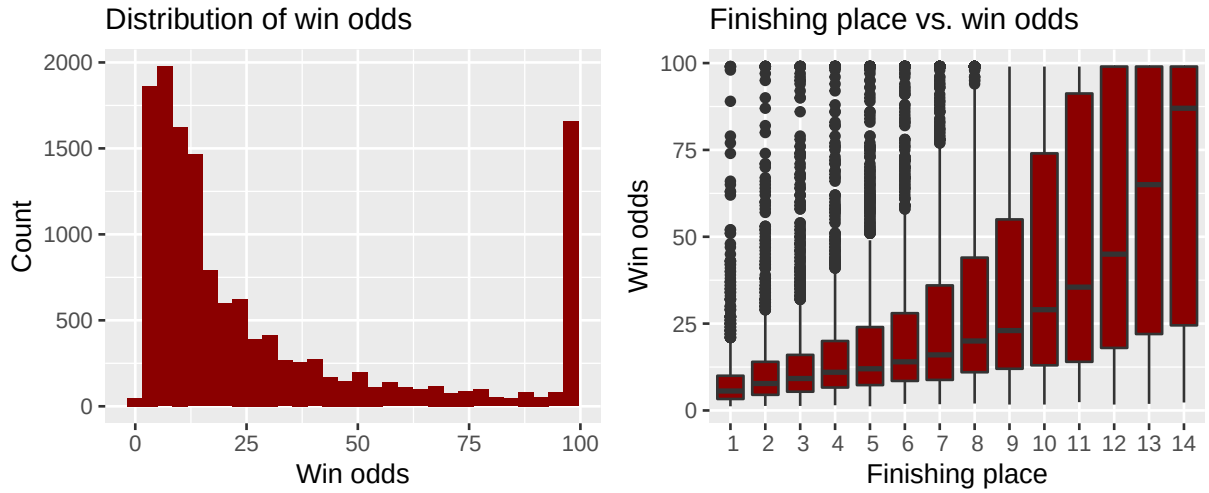


Moving towards the origin of the winning horses, looking at the graph, we can see that the country that produces the most winners is Australia. Interestingly, Australia exports their horses to most racing nations, especially those in the South-East area where half the horses in training come from Australia.

To see if biology had an effect, we took the winning horses and looked at their lineage. Looking at the left, we see that Exceed And Excel had children that racked up a total of 87 top 3 wins, and Starship had 16 top 3 wins. It makes sense that parents if the parents were successful racers, their offspring would be too.

	owner	Wins	Stake Total
1	Albert Hung Chao Hong	16.00	65045000.00
2	The Hon Ronald Arculli GBM GBS JP & Johanna K J Arculli BBS	16.00	61520000.00
3	Chan Ming Wing	15.00	42482000.00
4	Edmond Siu Kim Ping	15.00	85995000.00
5	Henry Cheng Kar Shun	14.00	41327000.00
6	Kwok Siu Ming	14.00	138312000.00
7	Martin Siu Kim Sun	14.00	79960000.00
8	David Hui Cheung Wing	13.00	56018000.00
9	Marces Lee Tze Bun	13.00	69670000.00
10	David Philip Boehm	12.00	131625000.00

To see if wealthier owners had an influence on horses winning, a comparison was made between the owner of the horse, the number of wins, and the average stake per race over the lifetime of the dataset. When putting a horse for a competition, owners must put in a stake, which comprises part of the prize money, the idea is that the stake is a proxy to wealth, the more a person/group has staked in their career the more money they have. The average staked amount in the data set $1,310,264. We see that all of the top ten winners spent more than average, indicating that wealth might be a factor in increasing the chances of winning



In Hong Kong decimal odds are used, which is the standard on most online betting sites. The number shows how much the total payout will be, including the original stake per unit staked. For example, a winning bet at 1.5 would return a total of 1.50 dollars for every 1 dollar staked, so the individual would have profited by 50 cents.

Starting with the graph on the left we see the win odds ranging from 1.2 to 99.0, and that most of the odds are on the lower end, with surprisingly a large count of odd of 99. On the right side, we again have the finishing place with the win odds, what we can see is that the most favorable horse has the lowest odds. Which makes sense as the payout should be less where there is more certainty a particular horse will win. What is also noticeable is that there are more outliers in the as you move to 14th place, which are instances where there has been an upset or an underdog has taken a top position.

Modeling

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 horse_{i1} + \beta_2 trainer_{i2} + \beta_3 actualwt_{i3} +$$

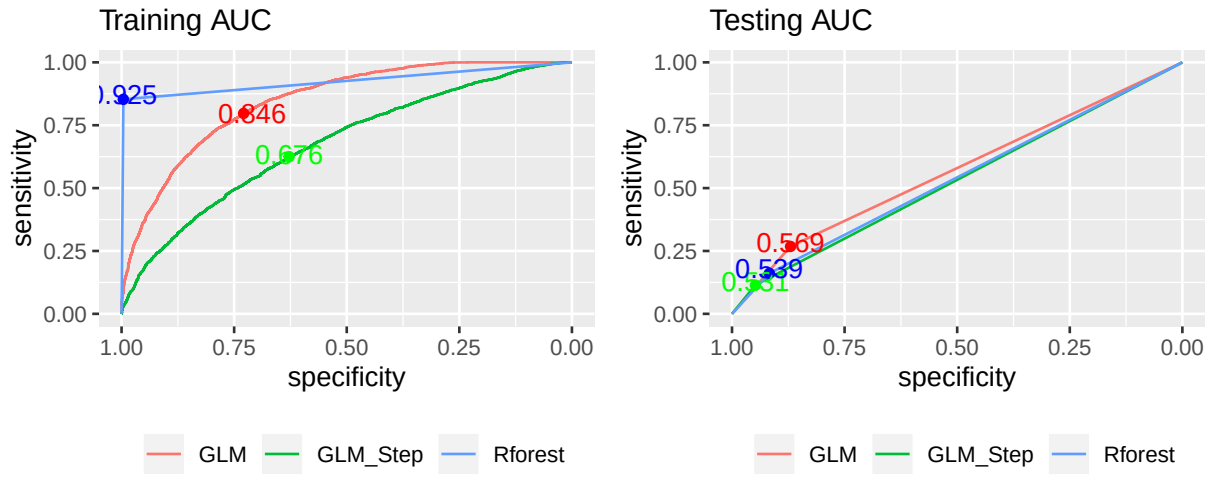
$$\beta_4 declarwt_{i4} + \beta_5 draw_{i5} + \beta_6 class_{i6} + \beta_7 distance_{i7} + \beta_8 course_{i8} + \beta_9 jockey_{i9}$$

Historically, Bill used a logistic regression, as such a logistic regression was used to predict whether or not a horse will win a given race. Looking at the formula, 9 variables were used as some variables are redundant, such as: dire, dam, country, etc. as those already describe the horse and are persistent over time.

	Coeff	Estimate	Std. Error	z value	Pr(>|z|)
194	'horseLUCKY BUBBLES'	8.19	1.59	5.15	0.00
275	'horseSMART GUY'	-5.20	1.55	-3.35	0.00
343	'jockeyJ Moreira'	1.43	0.37	3.83	0.00
370	draw12	-0.84	0.14	-5.99	0.00
354	'trainer.xJ Size'	3.57	0.92	3.86	0.00
362	declarwt	-0.01	0.00	-3.18	0.00
361	actualwt	-0.09	0.01	-13.58	0.00

Looking at a small subset of the summary results, one finding was that the likelihood of winning got worse as one was placed in the middle draw spots, and that being at the ends increases your chances of winning. Another finding was that, the horse most likely to win was Lucky Bubbles, while the horse most likely to lose was Smart Guy. Furthermore, only 2 jockeys were found to increase the chances of winning, which were

J Moreira and Z Purton. It also seems that weight, though signficant, has little to do with increasing the chances of winning.



Moving to predictions, rather than doing a random 80-20 train/test split, the test set contained the last race each horse ran, while the training data were all the races before in order to create a realistic scenario. For the models being used, the first was a logistic model that contained all the variables, the second model was a logistic "both" step model that used AIC, lastly was a random forest to see if another model could perform better than the logistic model. Referring to the graphs above rforest had the highest in sample AUC of 0.925, followed by the glm(0.846) and then the glm step(0.676). Unfortunatly, when running the logistic models, issues with perfect seperation occured. This is when the predicited values come as only 0 or 1, even after removing certain predictors the problem still persisted.

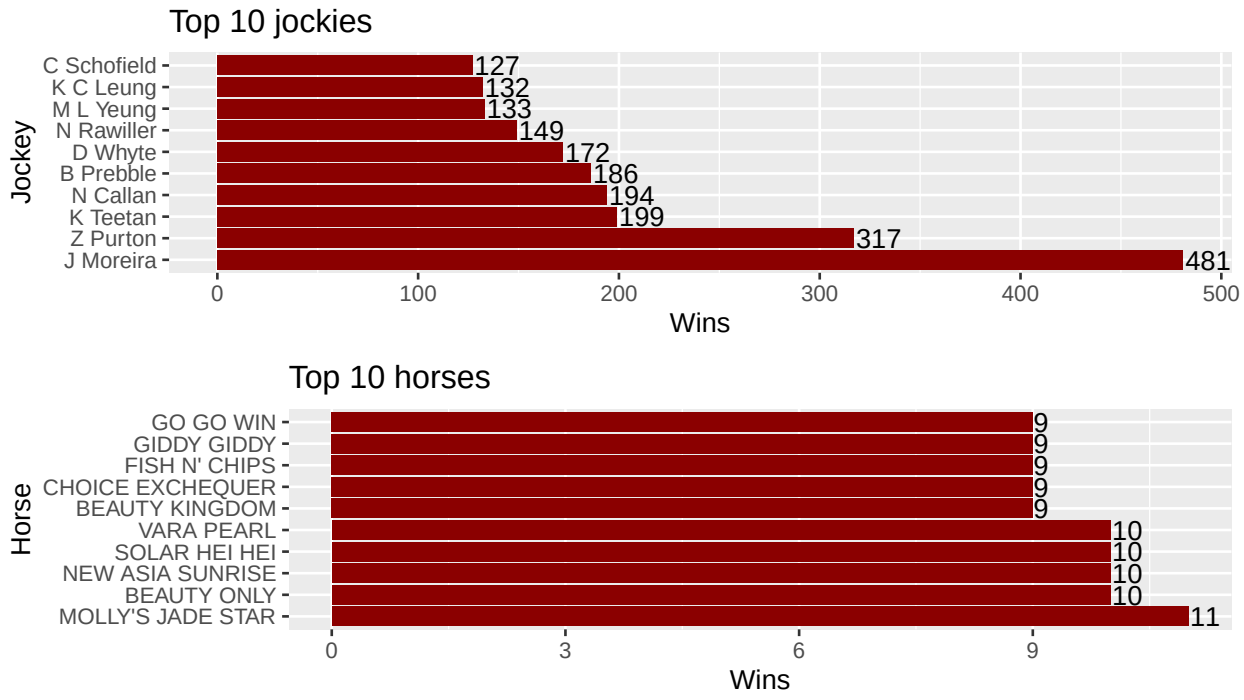
The sensitivities and specificities for rforest, glm, and glm step were (0.85,0.99), (0.80,0.73), and (0.62,0.63) all of which performed better than chance. However, once we applied these models to the test data set, we have a clear example of overfitting. The test AUC's for rforest, glm, and glm step were 0.539, 0.569, and 0.531 respectively. The sensitivities and specificities for the models were (0.16, 0.92), (0.27,0.87), and (0.11, 0.94). All the models were able to predict losing horses fine but are too poor in predicting winners.

Conclusion

Although we were not able to predict winning horses, we were able to see and understand what factors might increase/decrease the chances of a horse winning, as well as utilize some common machine learning methods. A major limitation was the lack of domain knowledge in the area of horse betting, some columns that required this domain knowledge such as "gear type" or "going" that would of helpful to understand. Another limitation was that the data spanned over 2 years, which is not enough of time to gather sufficient information on a horse's performance.

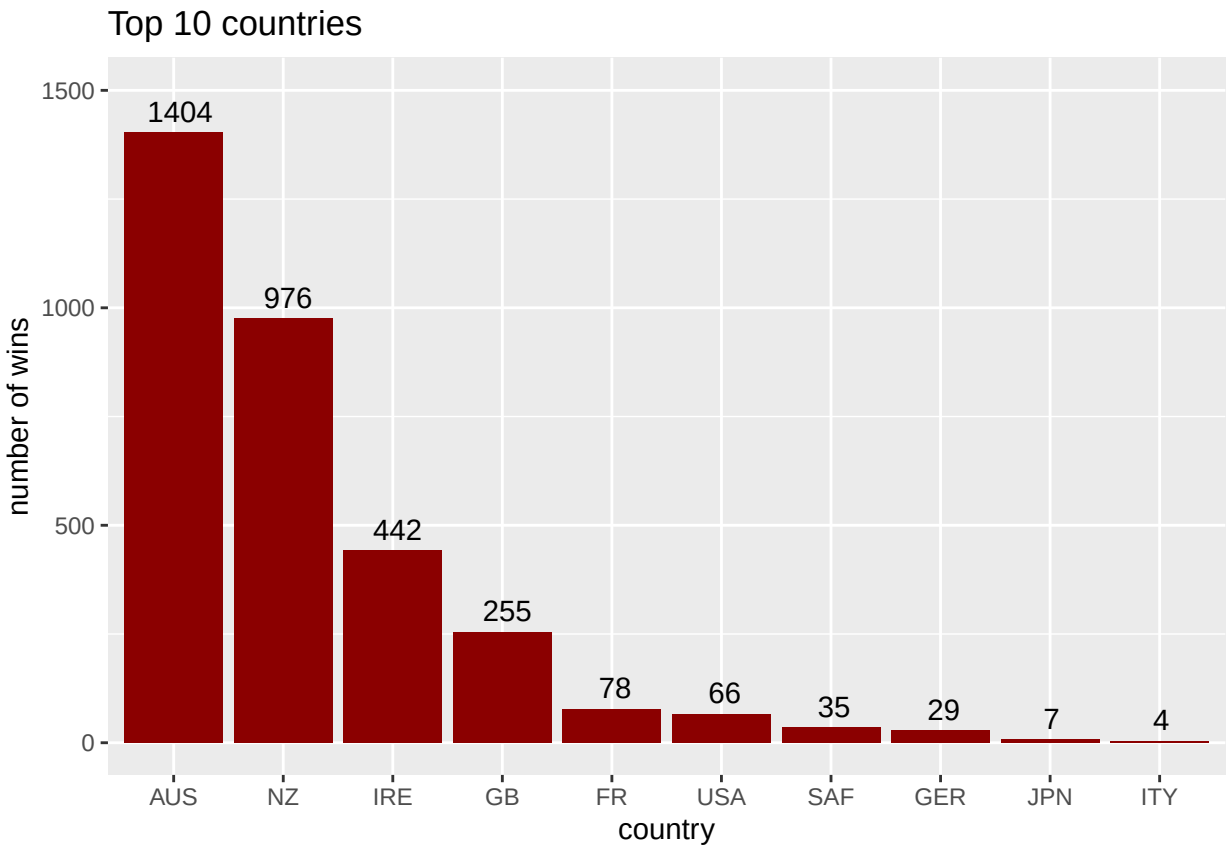
One way we can improve this experiment would be to include a dataset that would have the preliminary/training rounds of racing data and see how that impacts the current race. Another method that could be implemented is to use a forecast to predict the odds of horses winning, rather than using classification methods.

Appendix



To see who the best horses and jockeys were, a count was made everytime the horse or jockey was placed in the top 3 of races. Lookin at the jockies, J Moreira is the clear winner with 481 top 3 wins to his name, and Molly's Jade Star barely is the top 3 winner with only 11 wins. The reason why there is such a large gap of wins between the horses and the jockies is that jockies ride multiple horses during a racing event, as they get paid every time they race. The average fee a jocky gets is between 50 to 100 dollars per race, which explains the huge gap. In fact, jockies are the worst paid athletes with an average salary of 50,000 dollars per year.

```
### Country of origin ###
p7 = results %>%
  filter(is.na(country) != T) %>%
  group_by(country) %>%
  summarise(
    country_win =  sum(plc_top3, na.rm = T),
    count = n(),
  ) %>%
  arrange(desc(country_win)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(country, -country_win), y = country_win)) +
  geom_bar(stat = "identity", fill = "darkred") +
  ggtitle("Top 10 countries") +
  xlab("country") +
  ylab("number of wins") +
  coord_cartesian(ylim = 1:1500) +
  geom_text(aes(label= country_win), vjust = -0.5, hjust = .40)
p7
```


Top 10 countries
number of wins
1500
1000
500
0
1404
976
442
255
78
66
35
29
7
4
AUS
NZ
IRE
GB
FR
USA
SAF
GER
JPN
ITY
country