Milestone Report: Studying and Predicting Doping Related Incidents in New York's Horse Racing Industry

Springboard Capstone Project 1
Isabel van Zijl
November 28, 2017

The Problem

Horse racing is a sport that has been practiced since ancient times and today is a multi-billion dollar industry. It is a dangerous sport, both for the horse and the jockey, and has only grown to be more dangerous with the increased presence of doping abuse within the practice. Laws and regulations have been introduced to address safety and fairness concerns within the sport, but have they made a difference? Many trainers have been identified and punished over the years for doping their horses, but the problem seems to persist in the sport. Is there a way for us to predict which trainers are likely drugging their horses based on the histories of trainers that have already been suspended and/or fined for equine doping? If we can identify such trainers, what further legislation and regulations could be introduced to continue to build integrity within the sport?

Client

PETA (People for the Ethical Treatment of Animals) has always been a vocal supporter of animal rights and has taken action in the past to file complaints and influence policy regulation changes in the horseracing industry. In order to make convincing arguments about why certain policies should be changed, PETA needs as much evidence as possible to show probability of abuse. The more evidence they have, the more likely they can influence positive change for the welfare of racehorses. They can decide to lobby for stricter rules and regulations for the care of racing horses if provided with proper evidence. If I am able to identify a method of predicting what trainers are likely drugging their horses, investigation into and action against such trainers can be made.

Data Set

For this project, I'm using the Equine Death and Breakdown data set provided by the New York State Gaming Commission. This report lists horses that have broken down, been injured, or have died at race tracks in New York since March of 2009. The data set is downloadable as a CSV file and contains 3,240 records as of November 13, 2017. Since I'm also looking into whether doping is involved based on trainers histories, I built a smaller data set based on my own research that lists trainers that have either been suspended or fined for drugging one or more of their horses (titled Doping History Trainers data set). Article sources for each trainer are

provided in the csv file and the trainer names match exactly with the trainer names provided in the Equine Death and Breakdown data set. I attempted to find a data set that contained an official list of trainers that have a doping record, but I was unable to find one. For my project, I'm following the assumption that if a trainer has been fined or suspended for doping, it's likely that they probably drugged other horses that they trained in the past, which likely contributed to an incident taking place. There is no public record of the results of doping tests performed on specific horses, so I have to follow this assumption for my analysis. Under this assumption, we have to keep in mind that there could be more trainers and/or horses with a doping history but were not found in the research I performed.

Data Wrangling

The following steps were taken to clean the two data sets and then merge them:

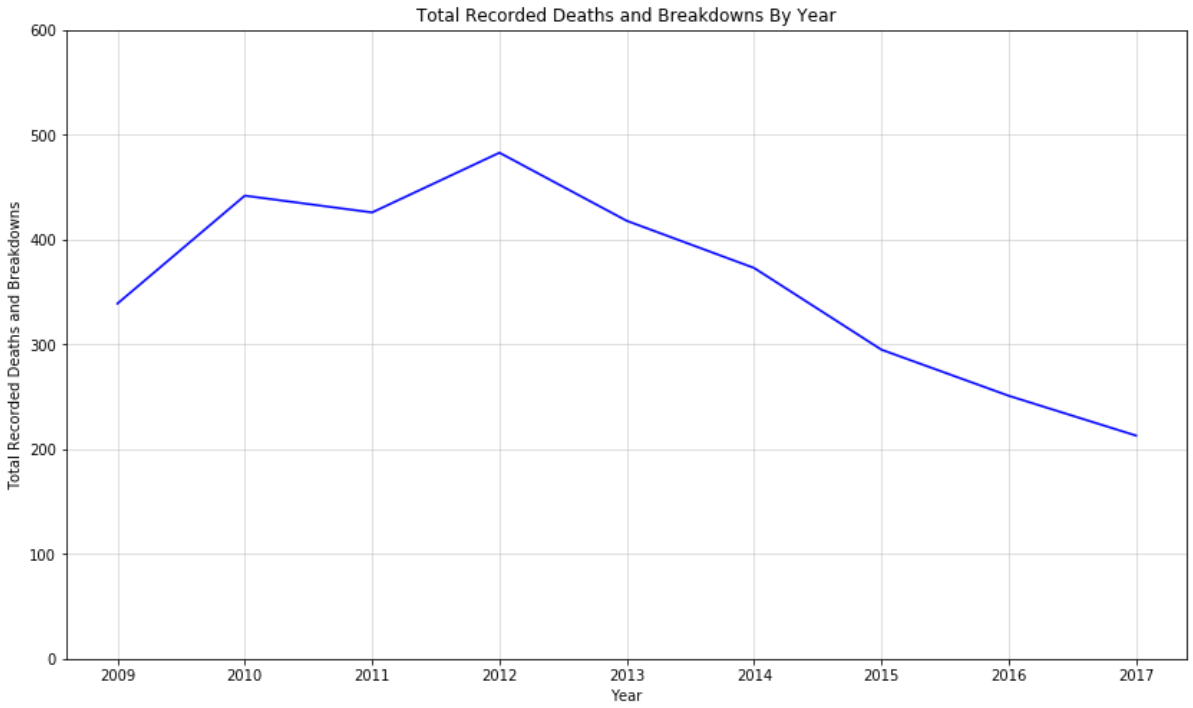
1. Identified inconsistent blank and null cells and converted them to NaN values.
2. Consolidated categories that had misspellings or were differently worded and then changed the datatype to categorical. Columns that were edited: *Incident Type, Track, Racing Type Description, Track*, and *Death or Injury* columns.
3. Create boolean weather condition columns based on the *Weather Conditions* descriptions. Eleven new columns were added based on the most common descriptors (e.g. Cloudy, Sunny, Etc.).
4. Merge (left join) the Equine Death and Breakdown data set with the Doping History Trainers data set on the trainer name (column name: *Trainer*).

Initial Findings

Most of the data in the data set is qualitative, but I was able to look at some trends over time and study which variables could have a large effect on yearly death and breakdown rates. The Equine Death and Breakdown dataset was first published in February 2013 but the oldest records in the set are from March 2009. The most recent version of the dataset contains information through November 2017. Although we haven't reached the end of 2017 yet, we can assume that we're close enough to the end of the year that there most likely won't be too much of an increase in the number of horses that will breakdown and/or die.

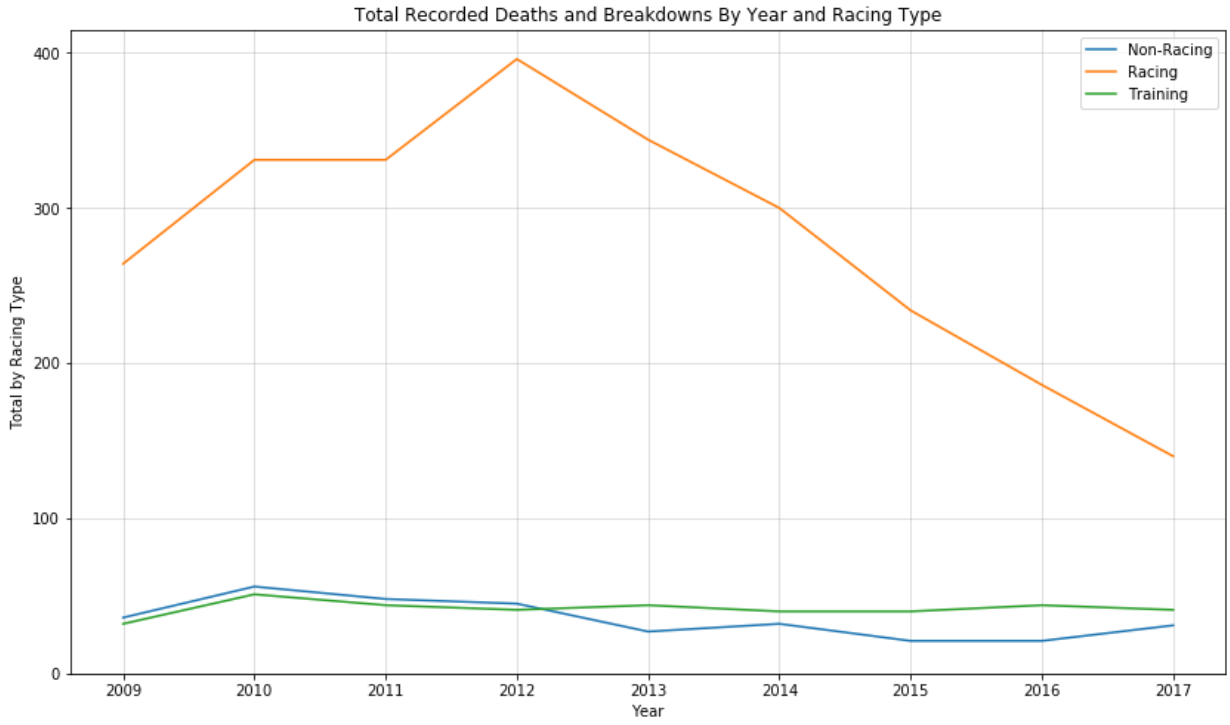
To start, I looked at the total number of deaths and breakdowns recorded per year (I will also refer to this as the yearly incident rate). In the figure below, we can see that total breakdowns and deaths have decreased over the years. The death and breakdown rate peaked in 2012 with a total of 483 deaths and breakdowns but has steadily been decreasing. The average of our totals is 360 incidents per year and the rate has stayed lower than that average since 2015. This means that the sport is headed in a good direction when it comes to protecting the the health and wellbeing of the horses since fewer horses are breaking down or dying on the track. This

might also signal that equine doping is decreasing in the industry, but we have to perform some deeper analysis to confirm this.



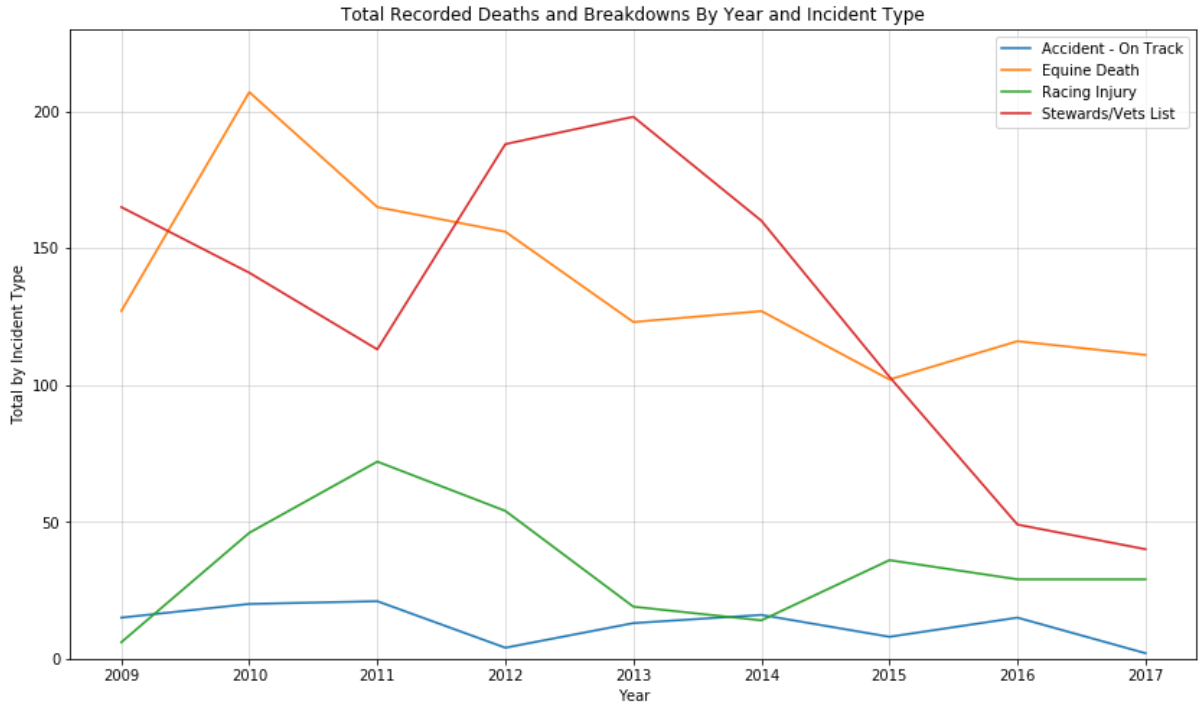
We were provided with three different *Racing Type* categories in our data set: racing, non-racing, and training. We can predict that the racing incidents will probably have a much higher incident rate than the other two racing types just by the nature of how races work. If trainers are going to drug their horses, it's usually going to be right before a race so the horse gets an adrenaline rush and can run its fastest while racing. For example, some of the trainers in our data set have been suspended for administering cocaine to their horses before racing.

In the figure below, we can see that we were correct in our predictions. Racing incidents occur at a high yearly rate while the number of training and non-racing incidents remain low and mostly constant over the years. Similar to what we saw in the previous figure for total deaths and breakdowns, there's a very steady drop in the number of racing incidents from 2012 onwards. This suggests that racing regulations could be helping to decrease the number of deaths and breakdowns that we're seeing on the racetrack, and therefore contributing to the overall decrease.



We should take note that the training and non-racing incident rates are probably artificially low. Racing incidents are easy to keep track of since races are watched by audiences and usually broadcasted. Training and non-racing incidents might just be self reported by owners and trainers, which could be why the totals are so low. This means that there actually could be more of a problem when it comes to horse deaths and breakdowns than our data suggests.

Another variable I wanted to study was *Incident Type*. There are eleven different categories in the incident type column, but I wanted to focus on four of the more common incident types for this analysis. In the figure below, we can see that accidents on the track have stayed relatively constant over the years, but there are some insights to be gained from the other three curves.

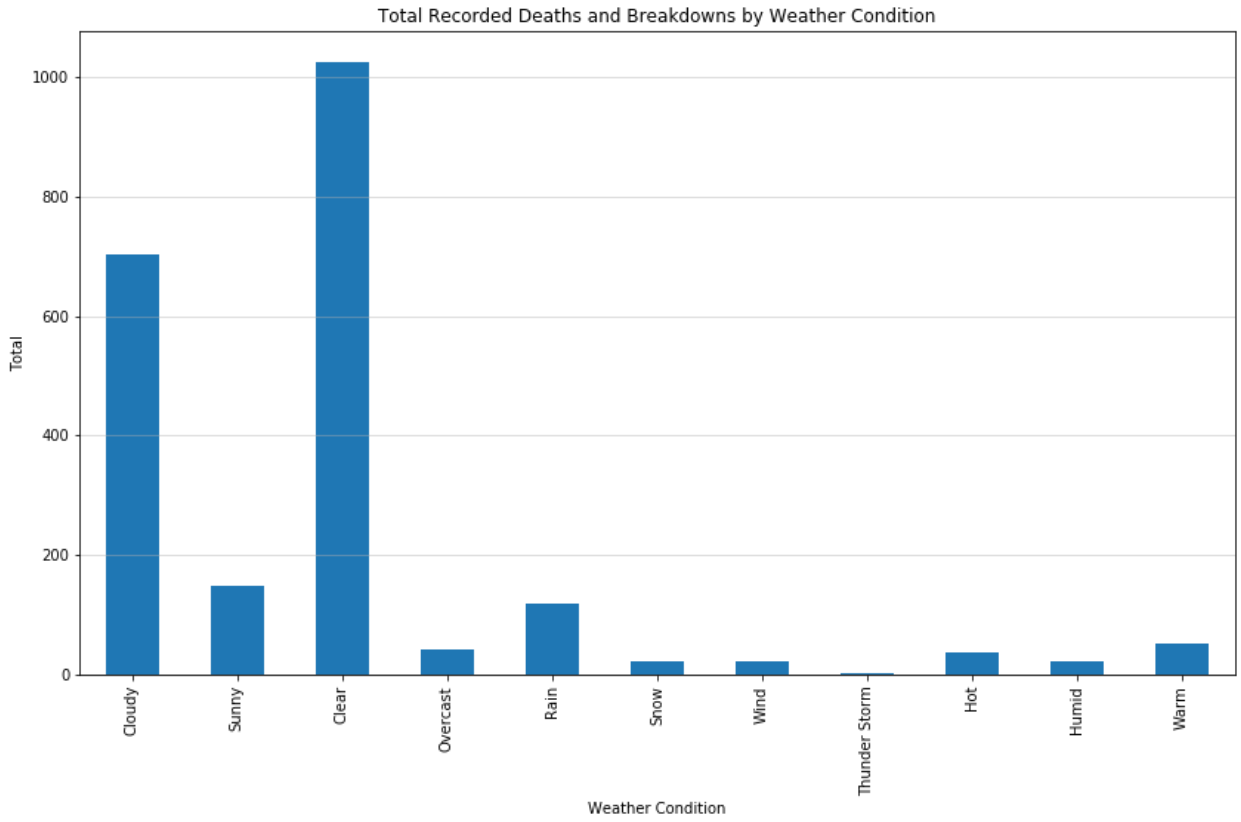


The Stewards List contains the names of horses that are ineligible race because of poor or inconsistent performance. They haven't been seriously injured, but are in bad enough shape that they've been deemed unraceable. In the above graph, we can see that the equine death and Stewards/vets List curves display a bit of an inverse relationship. Equine deaths peaked in 2010, but have slowly decreased over the years. Horses being placed on the Stewards/Vets List hit a low point in 2011, but increased in 2012, where it surpassed the number of equine deaths. We can also see that the number of racing injuries peaked in 2011 and has mostly decreased over the years.

Our graph could suggest that the number of deaths and racing injuries goes down when horses are placed on the Stewards/Vets List, as a precautionary move. Fewer horses have been placed on the Steward's List in recent years but the equine death rate remains steadily high. Perhaps Steward’s List requirements have gotten more lenient in recent years which is why we’re seeing this trend. What we can conclude from this graph is that perhaps more horses should be put on the Steward's List in order to try to reduce the number of equine deaths and injuries occurring each year.

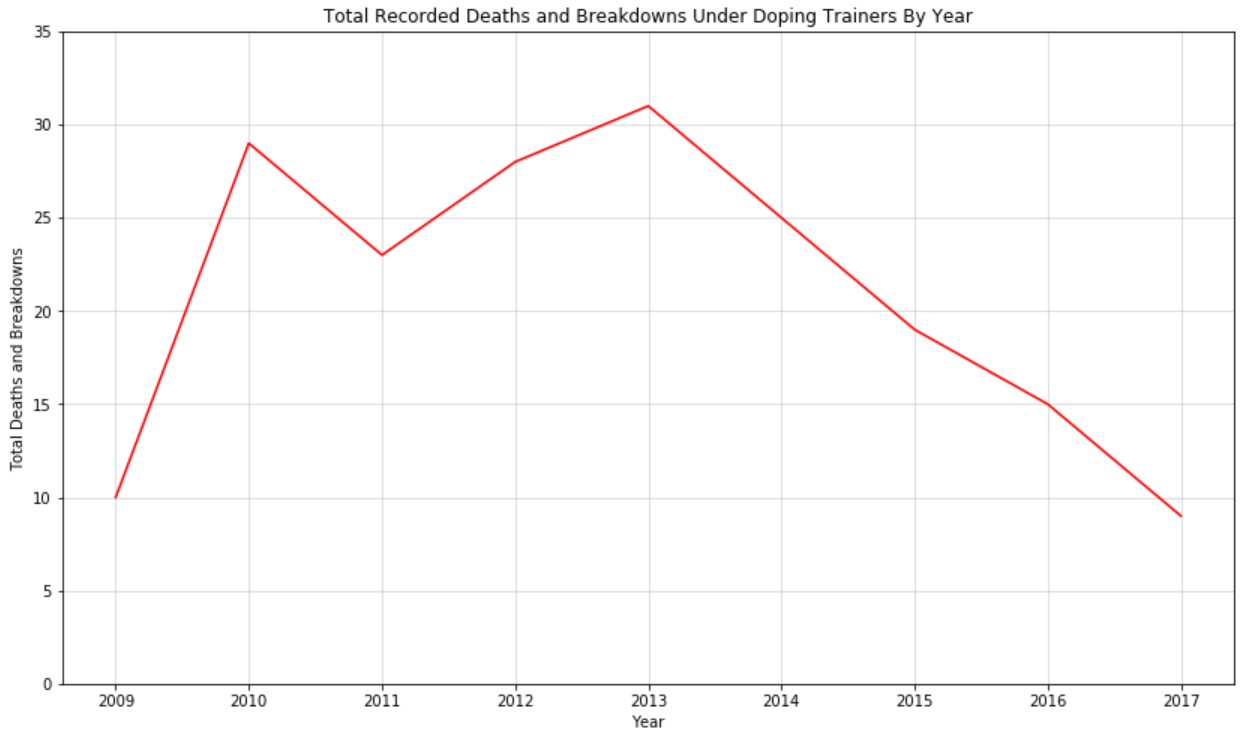
Something else to consider is that perhaps an impending breakdown or death isn’t noticeable until it’s too late, and therefore horses are never even placed on a Steward’s/Vets List to begin with. As mentioned before, some trainers will drug their horses right before the race in order to increase adrenaline levels. If a trainer decides to only drug their horse in a final round of racing, observation of an injury before the race could be highly unlikely. Adequate drug testing right before a race even begins could make a difference between an incident occurring or not.

Another variable that affects performance on the track is the weather. Sometimes it can be beneficial, and other times it can make or break a race. For example, if a track is too wet and muddy, it can be difficult for horses to run at quicker speeds and they end up over exerting themselves. It's also more likely that they could get a hoof stuck in the mud and collapse under such conditions. In our data set, we were provided with a *Weather Conditions* description column. I extracted some of the more common descriptions provided and built the following graph.



The most common weather condition recorded is by far the 'Clear' condition with over a thousand records. 'Cloudy' was also a relatively common weather condition recorded, followed by 'Sunny' and 'Rain'. What's surprising about these results is that the top three weather conditions don't necessarily yield poor track conditions. Let's say it's a clear and sunny day. Looking at the graph above, you would think that these conditions would have a high chance of yielding a breakdown or death. But the tracks themselves should be in relatively good shape under these conditions (assuming that the previous day's weather conditions didn't poorly affect them) and shouldn't have a large effect on racing conditions. We could also be seeing these results because races normally occur on days without extreme weather conditions anyway. From looking at this graph alone, it’s difficult to determine whether weather conditions are having a significant effect on yearly incident rates.

The main focus of this project is to study and predict likelihood of equine doping by looking at trainer histories. We're assuming that if a trainer has a history of doping (usually fined or suspended), it's likely that they've drugged many of their horses in the past. In the data set I built, there is a total of 22 trainers that have been suspended or fined for doping. Under these 22 trainers, there are 189 horses that have broken down or died on New York tracks since 2009. Some of our doping trainers have had a lot of horses that had incidents on the track (for example, Rudy Rodriguez has 50 horses recorded to have incidents), but the majority of these trainers only have a few horses that have had an incident. The following graph displays how these incidents occurred over time.



According to the above graph, equine deaths and breakdowns with a likelihood of doping peaked in 2013 and have decreased since then. This is a good sign that regulations have had an impact on both identifying trainers who drug their horses and preventing other trainers from taking up the practice. But yearly totals are still above zero as plenty of trainers have been identified as guilty of doping within the past year, so there is still work to be done.

Inferential Statistics

With the increasing prevalence of drug-abuse within the sport of horse racing over the past couple decades, the Thoroughbred Horseracing Integrity Act of 2015 was introduced to monitor and regulate drug usage and administration to American race horses under a national uniform standard. Effective on January 1, 2017, the legislation authorizes the Thoroughbred

Horseracing Anti-Doping Authority (THADA) to develop and administer a national anti-doping program. As an independent organization, THADA aims to uphold the integrity of the sport and focus on the welfare of the horses.

The Thoroughbred Horseracing Integrity Act has only been in effect for less than a year at this point, but we hope that the act has been able to make a difference in the sport already. We will be attempting to identify whether any differences in yearly death and breakdown rates are noticeable compared to our previous years of data. We will focus especially on horses with incidents under trainers with a history of doping.

As we discussed in the exploratory data analysis, the number of horses that have had incidents per year has decreased since 2012. At this point, we don't know if that decrease is statistically significant. The Thoroughbred Horseracing Integrity Act was enacted on January 1st, 2017, so we will test to see if the 2017 death and breakdown rate is significantly different from the mean of the other years. We only have 9 years worth of data total, so we will be performing a t-test. We do have to keep in mind that the year of 2017 is not yet complete, but we can assume that there most likely won't be many more incidents this year since it's November and therefore off season in New York.

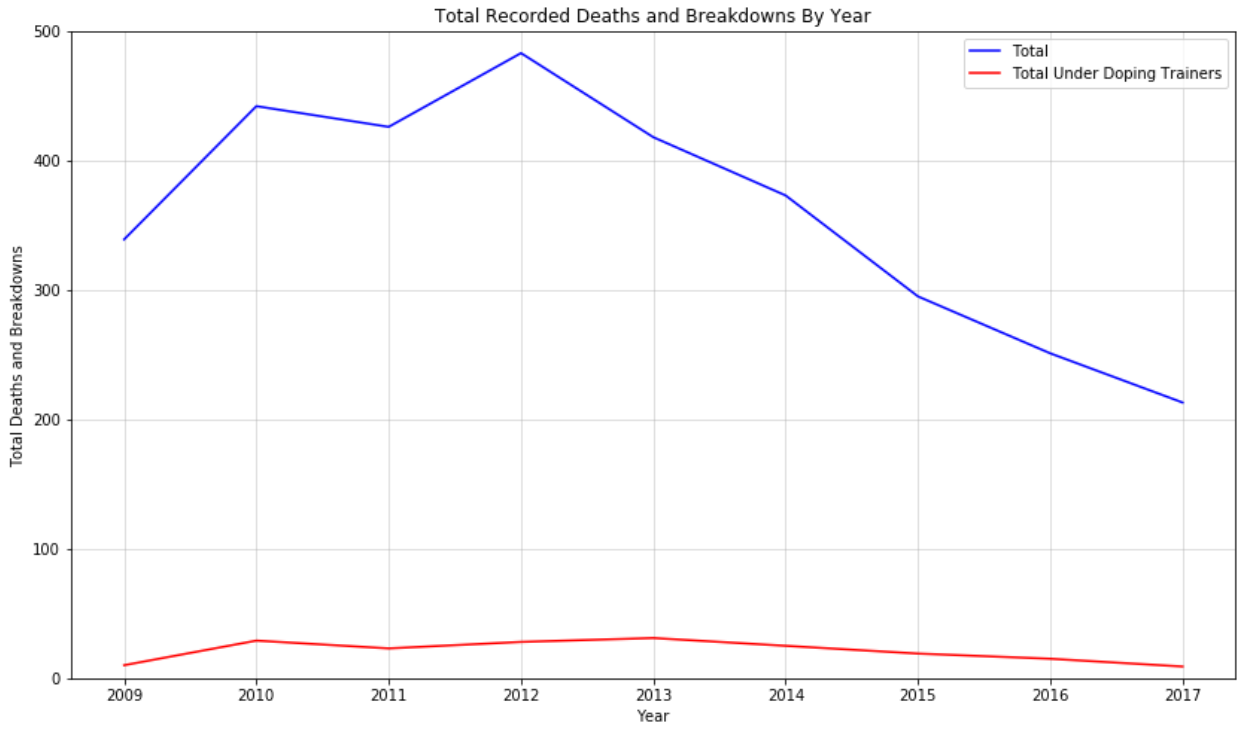
For this t-test, the null hypothesis that we're testing is that there's no significant difference in the 2017 death and breakdown rate and the average death and breakdown rate of other years (2009 - 2016). There were a total of 218 incidents in 2017 and an average total of 378 incidents per year in years 2009-2016. We performed the test and found a p value of essentially zero, revealing that we should reject the null hypothesis. There is indeed a statistically significant difference in the 2017 death and breakdown rate and the average death and breakdown rate of other years (2009 - 2016). This suggests that the sport really has been improving over the past few years, whether or not it's related to anti-doping legislation.

We do have to keep in mind that we are working with a small dataset. We only have 8 years worth of data (2009-2016) that we are using to calculate a pre-2017 mean. The fact that we still get a very small p value though suggests that the difference is significant.

Now that we know that there's been a significant decrease in the number of deaths and breakdowns per year, let's see if we can discover if it's related to the introduction of anti-doping programs. The null hypothesis for this t-test is that there is no significant difference in the 2017 death and breakdown rate and the average death and breakdown rate of other years (2009 - 2016) for horses trained by trainers with a history of doping. There were a total of 9 incidents in 2017 and an average total of 22.5 incidents per year in years 2009-2016 for horses trained by trainers with a history of doping. We calculated a p value of approximately 0.113, which does not pass the statistical significance test since it is greater than 0.55. Even if we were testing at the 10% level, our p value would still be too large. We then accept the null hypothesis in this situation. There is no significant difference in the 2017 death and breakdown rate and the average death and breakdown rate of other years (2009 - 2016) for horses trained by trainers

with a history of doping. If this is true, we have to wonder if legislation is actually making an impact on whether trainers are choosing to drug their horses.

To get a better idea of how these yearly rates look next to each other, I built this graph to see how the incident rates trended together over time.



We can see that there are much fewer incidents under trainers with a history of doping. We should keep in mind though that the total under doping trainers could be artificially low in the years 2009 to 2016. There wasn't a standard anti-doping test available yet and restrictions have gotten stricter over the years, so there could be unidentified trainers in that time span that regularly drugged their horses.

Since we've tested the incident rates separately, we should now test whether the proportional difference is significant. If the rate under doping trainers is decreasing simply because the overall rate is also decreasing, it doesn't really mean that anti-doping rules and regulations are making a difference.

Since we know that both the total yearly incident rate and the yearly incident rate under doping trainers have both been decreasing over the years, we will now test whether the change in the proportion of drugged horses out of the total is significant. We will be comparing two values, so we will perform another t-test. The null hypothesis for this test is that there is no significant difference in the 2017 proportion of incidents under doping trainers to total incidents and the average proportion of incidents under doping trainers to total incidents of other years (2009 -

2016). The 2017 proportion of doping trainers to non-doping trainers was 4.225 and the average proportion of doping trainers to non-doping trainers for years 2009-2016 was 5.905. The difference in these two values seems small, but we end up with a p value of 0.007, proving that the difference is statistically significant and that we should reject the null hypothesis. This is a good sign that doping practices are actually decreasing since we're looking at a percentage of a total. The fact that yearly incident rates themselves are decreasing and yearly incident rates under doping trainers are decreasing even more so signals that the sport is improving its practices in the state of New York.

Milestone Report Conclusions

Though yearly equine death and breakdown rates have been decreasing, there is always room for improvement. Most of our records are racing related breaking downs or deaths, and fortunately the yearly totals of these racing incidents have decreased over the years. We saw that adding horses to the Stewards/vets list most likely helped save some horses from worse injuries and possibly death, but the rate of an incident occurring (especially death) is still quite high in 2017 compared to other incident types. We also saw that a weather condition variable likely can't singularly reveal whether an incident is likely.

In our exploratory data analysis, we found that the yearly incident rate under doping trainers has been decreasing, but we didn't know if this decrease was significant. Though we didn't find the decrease of the 2017 incident rate under doping trainers compared to other years to be significant in our t-test, we still identified statistical significance in our two other tests. The significant decrease in the proportion of incidents under doping trainers in 2017 compared to the average proportion of incidents under doping trainers for the previous eight years suggests that legislation and regulations could be making a difference in whether trainers choose to drug their horses.

Though the number of horses under doping trainers has decreased over the years, it doesn't mean that the practice of doping is coming to an end. The mindset of cheating the system and winning more races (and money) will always be present. New drugs are always being developed and drug testing is only administered at races, so there are definitely some gaps in the above analysis for non-racing and training incidents. With the development of more robust anti-doping tests and more stringent rules, hopefully the number of deaths and breakdowns per year will continue to decrease and the racetrack can return to being a level playing field for all participants.