Homework 2

Joyce LJ Woznica

# Introduction

Each race conducted at a New York Thoroughbred racetrack is observed by three stewards and at the harness tracks, each race is observed by three. The stewards’ and judges’ viewing stand is located near the finish line of each racetrack and is equipped with several television monitors to permit the viewing of multiple angles of each race. The stewards and judges observe the races and the race grounds to ensure that all conduct is in accordance with rules and regulations.

If individuals that break rules are cited, this encourages encourage compliance and mitigates risk. The horse racing industry is plagued with accidents ranging from jockey or driver injuries to equine death – rules exist to attempt to minimize these incidents. Further investigation can shed light on contributing factors to broken rules and associated incidents and provide insight into what leads to the high number of infractions. Armed with this information, tracks can put the proper constraints in place in the future to minimize infractions that could also minimize incidents.

# Data

To begin such an investigation, a subset of both harness and thoroughbred racetracks within New York State (NYS) will be studied. For this assignment, I used following two sources – the data set detailed below as well as tweets from various twitter accounts that follow the race industry. It was approved through email that I was able to use the data in these sets for semi-structured data despite the fact that they are not in JSON or social media.

## New York State Race Track Rulings from 1985 to Present

There are few racing jurisdictions that have equaled New York’s strong stand on integrity in horse racing. Over the past decade, the Board has enabled more capable and qualified people to become stewards and presiding judges at racing events. Rulings are captured and detailed in this data set.

| **Data Label** | **Data Type** | **Data Description** |
| --- | --- | --- |
| Fine Year | Numeric | Year of ruling |
| Notice | Text | Notice ID number |
| Full Name | Text | Licensee’s full name |
| Race Track | Text | Issuing location |
| Race Type | Text | Racing division: Harness or Thoroughbred; “?” indicates unknown |
| Type | Text | Type of ruling: Fine, Suspension, Denial |
| Notice Date | Date | Date of Notice for ruling |
| Rules | Text | Rules violated; public reference list of rules available at http://www.gaming.ny.gov/statutes.php |
| Ruling Text | Text | Text description of the ruling |
| Occupation | Text | Licensee’s occupation |
| Fine Amount | Numeric | Fine amount in dollars |
| Suspension Start Date | Date | Suspension starting date |
| Suspension End Date | Date | Suspension ending date |
| Days Suspended | Text | Number of days suspended or description of period of suspension |

Source: <https://data.ny.gov/Government-Finance/Horse-Racing-Rulings-Beginning-1985/igam-2tkj>

## Twitter Accounts/Hashtags

To gather additional information about how the race industry is seen from the eyes of observers as well as what it made public about certain infractions or incidents, the following twitter accounts and hashtags were reviewed.

| **Twitter Account or Hashtag** | **Data Description** |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Data Preparation Plan

A subset of the data will used for this assignment to only contain data for the years of 2009 – 2019. Data may be removed or manipulated to obtain the proper results and/or graphs for analysis purposes.

All data will be read into a *pandas* data frame for manipulation. Outputs with summarized or gathered information from the source data was also placed into a data frame before plotting with *seaborn*.

The code will be well documented and will be separated into the following sections:

* Package Import  
  In this section, the packages that will be used as part of the program are imported.
* Data Load, Structure and Preparation  
  In this section, the data is loaded by reading the *csv* file into a *pandas* dataframe for manipulation. A full path name was provided to the datafile for reading. As mentioned, earlier years were removed from the data so the data only reflects the last decade. Missing data was investigated, but there was no need to remove missing data from the data set.
* Merging of Data Sets  
  In order to review the possible incidents and correlation to any rules broken on the same date on the same track, the two data sets will have to be merged on these values:

|  |  |
| --- | --- |
| **Rulings Data Set Data Label** | **Incidents Data Set Data Label** |
| Notice Date | Incident Date |
| Race Track | Track |

Additional merging may be required.

* Analysis and Graphing  
  The next sections in the program consists of manipulating the data into different subsets to answer specific questions and for graphing purposes. The *seaborn* package was used for graphing.

## Overall Data Approach

A subset of the data was used for this assignment to only contain data from the years of 2009 – 2019 so that this dataset could be used in conjunction with the other data sets which only has data from the year starting in 2009. This reduced the overall dataset to 11,516 rows and 14 columns.

No data was removed because when evaluated, the missing data was not required and therefore would not affect the data results.

All data was read into a *pandas* data frame for manipulation. Outputs with summarized or gathered information from the source data was also placed into a data frame before output into file formats.

# Analysis

As part of reviewing this data set, the following information is to be analyzed.

## Infractions

### Common Infraction Words

Question 1: Are their common threads or words in the descriptions of the infraction?

![A close up of a piece of paper

Description automatically generated]()

## Relationship to Incidents and Infractions

### Common Infraction/Incident Words

Question 2: Are there common threads or words in the descriptions of the infraction as related to a possible incident on the same day at the same track?

### Correlation with Incidents and Infractions

Question 3: Is there any correlation between incidents and recent infractions?

# Conclusion