Big Data Derby 2022 Kaggle Challenge

Enrollment. No.(s) - 9919103037, 9919103057, 9919103119
Name of Student (s) – Manas Dalakoti, Shubham Garg, Gaurav Kumar
Name of supervisor – Dr. Shikha Mehta

Department of CSE/IT

Jaypee Institute of Information Technology University, Noida

October 2022

Submitted in partial fulfillment of the Degree of Bachelor of Technology

in

Computer Science Engineering

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING & INFORMATION TECHNOLOGY

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

TABLE OF CONTENTS

Chapter No. Topics **PageNo.**

1. Introduction

1.1 General Introduction

Injury prevention is a critical component in modern athletics. Sports that involve animals, such as horse racing, are no different than human sport. Typically, efficiency in movement correlates to both improvements in performance and injury prevention.

A wealth of data is now collected, including measures for heart rate, EKG, longitudinal movement, dorsal/ventral movement, medial/lateral deviation, total power and total landing vibration. The data science skills and analysis are needed to decipher what makes the most positive impact.

In this competition, it is expected to create a model to interpret one aspect of this new data. The contributors were among the first to access X/Y coordinate mapping of horses during races. Using the data, they might analyze jockey decision making, compare race surfaces, or measure the relative importance of drafting. With considerable data, contestants can flex their creativity and problem solving skills.

The New York Racing Association (NYRA) and the New York Thoroughbred Horsemen's Association (NYTHA) conduct world class thoroughbred racing at Aqueduct Racetrack, Belmont Park and Saratoga Race Course.

With their help, NYRA and NYTHA will better understand their vast data set, which could lead to new ways of racing and training in a highly traditional industry. With improved use of horse tracking data, this analysis could help improve equine welfare, performance and rider decision making.

BIG DATA Derby

1.2 Problem Statement

The problem statement is to generate actionable, practical, and novel insights from horse tracking data that devises innovative and data-driven approaches to analyzing racing tactics, drafting strategies and path efficiency. There are several potential topics to analyze.

These include, but are not limited to:

- Create a horse rating measuring expected finish position versus actual finish position. How does a horse's expected finish position change through the running of a race? Does this rely on a horse's own position or is it influenced by the position of competitors?
- What are optimal racing strategies? Considering different venues, surfaces and race distances. Create a jockey rating based upon path efficiency?
- Create a surface measure model which would rate the fairness of different paths on a racecourse that may be beneficial or harmful to finish position based. This may be a result of unknown barometric, weather or maintenance factors.
- Create a model that reveals optimal gait patterns. Does the model differ for such factors as age, distance, race section or surface?

1.3 Significance/Novelty of the problem

The significance of this work is to analyze horse racing tactics, drafting strategies, and path efficiency and to develop a model using never-before-released coordinate data along with basic race information.

This work will help racing horse owners, trainers, and veterinarians better understand how equine performance and welfare fit together. With better data analysis, equine welfare could significantly improve.

1.4 Empirical Study

Our sport is currently investing significant money in collecting far more precise tracking data in the hopes of improving equine welfare. Along with stride data, we can now collect measures for heart rate, EKG, longitudinal movement, dorsal/ventral movement, medial/lateral deviation, total power and total landing vibration. However, we do not have analysts with the appropriate expertise to help decipher these data sets.

We hope this competition allows us to interact with data scientists to help find solutions to equine safety issues as well as develop a roster of academics and motivated hobbyists who lead us in analyzing the coming generations of data.

1.5 Brief Introduction of Solution approach

We aim at performing Data pre-processing and feature engineering to prepare the dataset.

The task is to:

- Perform initial data exploration.
- Perform data pre-processing.
- Perform EDA in a cleaned data set.
- Perform feature engineering.
- Perform time series forecasting.

Chapter-2 Literature Survey

2.1 Summary of papers studied

Paper1:https://www.researchgate.net/publication/50946368_Exploratory_data_analysis_in_the_context_of_data_mining_and_resampling

This paper is based on the fact that the conventional conceptual frameworks of EDA might no longer be capable of coping with this trend. EDA is introduced in the context of data mining and resampling with an emphasis on three goals: cluster detection, variable selection, and pattern recognition. Two Step clustering, classification trees, and neural networks, which are powerful techniques to accomplish the preceding goals, respectively, are illustrated with concrete examples. This article introduces several new EDA tools, including TwoStep clustering, recursive classification trees, and neural networks, in the context of data mining and resampling, but these are just a fraction of the plethora of exploratory data mining tools. In each category of EDA there are different methods to accomplish the same goal, and each method has numerous options (e.g. the number of k-fold cross-validation). Last but not least, exploratory data mining techniques could be simultaneously or sequentially employed. For example, because both neural networks and classification trees are capable of selecting important predictors, they could be run side by side and evaluated by classification agreement and ROC curves. On other occasions, a sequential approach might be more appropriate. For instance, if the researcher suspects that the observations are too heterogeneous to form a single population clustering, could be conducted to divide the sample into sub-samples. Next, variable selection procedures could be run to narrow down the predictor list for each sub-sample. Last, the researcher could focus on the inter-relationships among just a few variables using pattern recognition methods. The combinations and possibilities are virtually limitless. Data detectives are encouraged to explore the data with skepticism and openness.

Paper2: https://www.researchgate.net/publication/308007227_Exploratory_Data_Analysis

Exploratory data analysis (EDA) is an essential step in any research analysis. The primary aim with exploratory analysis is to examine the data for distribution, outliers and anomalies to direct specific testing of your hypothesis. It also provides tools for hypothesis generation by visualizing and understanding the data usually through graphical representation . EDA aims to assist the natural patterns recognition of the analyst. Finally, feature selection techniques often fall into EDA. Since the seminal work of Turkey in 1977, EDA has gained a large following as the gold standard methodology to analyze a data set . According to Howard Seltman (Carnegie Mellon University), “loosely speaking, any method of looking at data that does not include formal statistical modeling and inference falls under the term exploratory data analysis”. EDA is a fundamental early step after data collection and preprocessing , where the data is simply visualized, plotted, manipulated, without any assumptions, in order to help assess the quality of the data and build models. “Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to explore, and graphics gives the analysts unparalleled power to do so, while being ready to gain insight into the data. There are many ways to categorize the many EDA techniques”.

Paper3:https://www.researchgate.net/publication/359182169_Optimal_Model_of_Horse_Racing_Competition_Decision_Management_Based_on_Association_Rules_and_Neural_Network

Horse racing is also a worldwide traditional competition event. There are 3.2 million competitions each year, and the total prize money for horse racing in the world amounts to 360 billion yuan. Information technology has played a huge role in promoting the development of horse racing. i.e. advancement of science and technology has not only promoted the advancement of horse racing but also has higher requirements for the decision-making management and performance prediction of horse racing. Based on the research of a large amount of data, neural network algorithms establishes a horse racing competition decision-making management optimization model. It provides decision-makers with powerful means and tools, provides an effective quantitative basis for making reasonable training management decisions and training programs, and provides scientific predictions for public. Among the participants of high-level equestrian competitions, the injury rate is also the highest, and equestrian competitions require skilled emergency medical services. Many factors in horse racing have a direct impact on the outcome of the race. To construct a horse racing decision-making model, the most important thing is to understand the factors that affect horse performance, including gender, scoring, ranking, weight, age, horse top three rate, harness, race schedule, venue, nature of the venue, and jockey.

2.2 Integrated summary of the literature studied

Horse racing is a speed competition that tests humans to control and control horses. It is one of the main events of equestrian sports. Its forms are changeable, but the principles are basically speed competitions. Horse racing is also a worldwide traditional competition event. According to textual research, horse riders were found in stone carvings of the Neolithic Age, and as early as the beginning of the seventh century BC, the four-horse driving sports event appeared in the Olympic Games held in ancient Greece. Later, the project was changed to human driving for competition. In modern times, with the continuous development and prevalence of horse racing, organization, management, and competition have become more and more important compared with ancient horse racing events, and the method is more scientific and more advanced. With the continuous popularity of horse racing, many scholars have conducted research on horse racing. Ghezelsefloo proposed: based on the theory of service operation management, transplant the theoretical knowledge of the discipline system of service operation management into the operation management of horse racing events. Zhang and Liu proposed horses are a unique match between two life forms of equestrian events, and the health of horses directly affects the event, which requires a very high level of organization and management of our events. Research by Quintana et al. shows that among the participants of high-level equestrian competitions, the injury rate is also the highest, and equestrian competitions require skilled emergency medical services. Research by Fenner et al. pointed out that it is necessary to establish a comprehensive long-term planning system for events, pursue the irreplaceability of the equestrian event brand, develop the characteristic highlights of the equestrian event, and propose the establishment of a professional talent training system. Research by Sun and Li shows that in the context of equestrian sports, competitions can promote and drive related consumption. Padalino et al. explored the characteristics of the development of large-scale sports events in Hong Kong, analysis of sports development and management mechanisms, and the characteristics of horse racing to drive and promote the development of Hong Kong’s gaming industry and sports culture. Hoseini and Amani proposed an equestrian event management system based on the B/S model to facilitate event informatization. Information technology is the foundation and symbol of modern society. With the development of artificial intelligence and big data, many experts use information tech- nology to manage and predict sports events. At present, the research and application of predictive management have penetrated into various research fields, and the field of horse racing research is no exception. Predictions can be made in horse racing strategy research, decision-making, coaches, athlete team construction, training programs, and so on. In recent years, different prediction algorithms have been widely used in nonlinear prediction in various fields, such as BP neural network, wavelet neural network, support vector machine regression, and so on. Among them, association rules and neural networks have become the most successful prediction methods in the field of horse racing because they are easy implement any complex nonlinear mapping function. Forecast through scientific methods; recognize the direction, trend, and law of the development and change of things; and take effective measures to control the codevelopment, so as to better optimize the management of horse racing decisions.

Chapter 3: Requirement Analysis and Solution Approach

3.1 Solution Approach

The problem requires us to understand the patterns and visualize them from the given dataset. This all needs to be done to improve the health of horses that participate in the derby each year as well as introduce new safety measures and understand the future of the horse using time-series-forecasting.

We would need some tools and frameworks that can handle the huge amount of data provided.

For initial analysis, performed on 20% of the dataset, we use python programming language and Google Colab for the memory requirements.

We perform the univariate and bivariate analysis of all the variables in the given data files using data modeling and visualization libraries such as pandas for dataframe management, matplotlib/seaborn for all the visualizations. The google colabs's powerful compute engine allows us to perform all the above tasks.

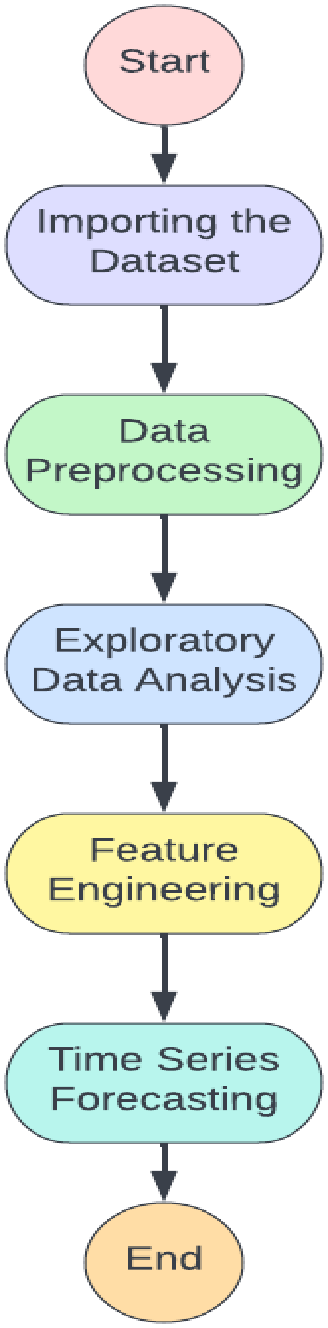
For the huge dataset as a whole to gain complete insights, we would use big data frameworks and warehouses such as bigquery to store all the given dataset. Since the data is already cleaned, we can perform analysis directly using any data manipulation tools such as dbt, snowflake for bigquery itself.

We would be using SQL for all the analysis tasks and some visualization libraries such as d3.js or chart.js integrated into a webapp.

Chapter-4 Modeling and Implementation Details

4.1 Implementation details:

The following flow chart shows the implementation that we would perform on the initial 20% dataset to understand the patterns better using python programming language.



Dataset Description:

File descriptions

- nyra_start_table.csv - horse/jockey race data
- nyra_race_table.csv - racetrack race data
- nyra_tracking_table.csv - tracking data
- nyra_2019_complete.csv - combined table of three above files

Columns

nyra_start_table.csv

- track_id - 3 character id for the track the race took place at. AQU -Aqueduct, BEL - Belmont, SAR - Saratoga.
- race_date - date the race took place. YYYY-MM-DD.
- race_number - Number of the race. Passed as 3 characters but can be cast or converted to int for this data set.
- program_number - Program number of the horse in the race passed as 3 characters. Should remain 3 characters as it isn't limited to just numbers. Is essentially the unique identifier of the horse in the race.
- weight_carried - An integer of the weight carried by the horse in the race.
- jockey - Name of the jockey on the horse in the race. 50 character max.
- odds - Odds to win the race passed as an integer. Divide by 100 to derive the odds to 1. Example - 1280 would be 12.8-1.
- position_at_finish - An integer of the horse's finishing position. (added to the dataset 9/8/22)

nyra_race_table.csv

- track_id - 3 character id for the track the race took place at. AQU -Aqueduct, BEL - Belmont, SAR - Saratoga.
- race_date - date the race took place. YYYY-MM-DD.
- race_number - Number of the race. Passed as 3 characters but can be cast or converted to int for this data set.
- distance_id - Distance of the race in furlongs passed as an integer. Example - 600 would be 6 furlongs.
- course_type - The course the race was run over passed as one character. M - Hurdle, D - Dirt, O - Outer turf, I - Inner turf, T - turf.
- track_condition - The condition of the course the race was run on passed as three characters. YL - Yielding, FM - Firm, SY - Sloppy, GD - Good, FT - Fast, MY - Muddy, SF - Soft.
- run_up_distance - Distance in feet of the gate to the start of the race passed as an integer.

- race_type - The classification of the race passed as as five characters. STK - Stakes, WCL - Waiver Claiming, WMC - Waiver Maiden Claiming, SST - Starter Stakes, SHP - Starter Handicap, CLM - Claiming, STR - Starter Allowance, AOC - Allowance Optionl Claimer, SOC - Starter Optional Claimer, MCL - Maiden Claiming, ALW - Allowance, MSW - Maiden Special Weight.
- purse - Purse in US dollars of the race passed as an money with two decimal places.
- post_time - Time of day the race began passed as 5 character. Example - 01220 would be 12:20.

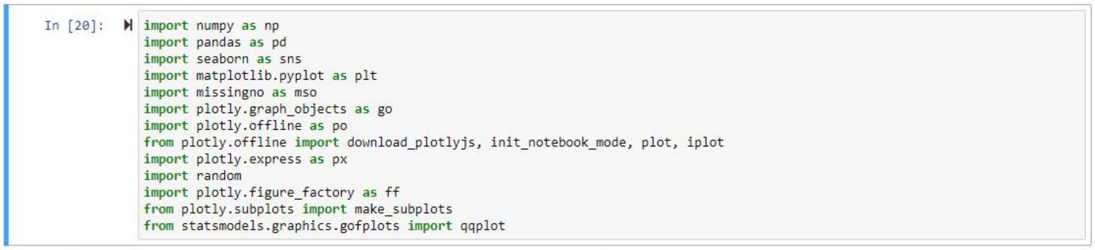
nyra_tracking_table.csv

- track_id - 3 character id for the track the race took place at. AQU -Aqueduct, BEL - Belmont, SAR - Saratoga.
- race_date - date the race took place. YYYY-MM-DD.
- race_number - Number of the race. Passed as 3 characters but can be cast or converted to int for this data set.
- program_number - Program number of the horse in the race passed as 3 characters. Should remain 3 characters as it isn't limited to just numbers. Is essentially the unique identifier of the horse in the race.
- trakus_index - The common collection of point of the lat / long of the horse in the race passed as an integer. From what we can tell, it's collected every 0.25 seconds.
- latitude - The latitude of the horse in the race passed as a float.
- longitude - The longitude of the horse in the race passed as a float.

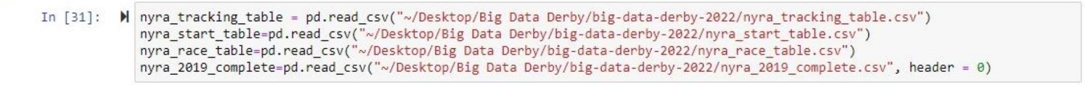
nyra_2019_complete.csv - This file is the combined 3 files into one table. The keys to join them trakus with race - track_id, race_date, race_number. To join trakus with start - track_id, race_date, race_number, program_number.

- track_id - char(3)
- race_date - date
- race_number - char(3)
- program_number - char(3)
- trakus_index - int
- latitude - float
- longitude - float
- distance_id - int
- course_type - char(1)
- track_condition - char(3)
- run_up_distance - int
- race_type - char(5)
- post_time - char(5)
- weight_carried - int
- jockey - char(50)
- odds - int
- position_at_finish - An integer of the horse's finishing position. (added to the dataset 9/8/22)

Importing the Dataset

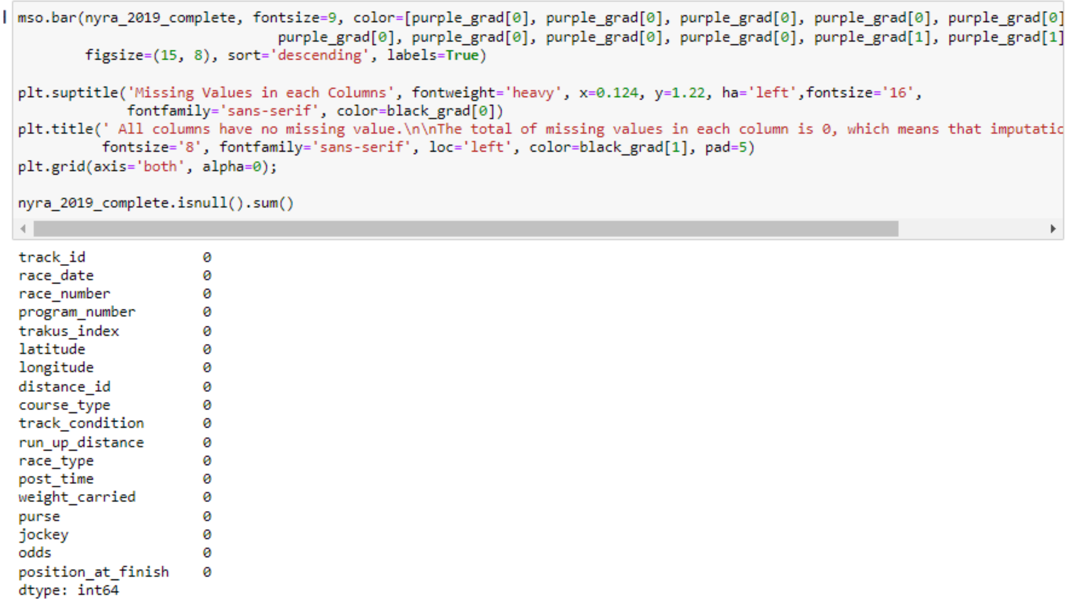


```
In [20]:  import numpy as np
          import pandas as pd
          import seaborn as sns
          import matplotlib.pyplot as plt
          import missingno as mso
          import plotly.graph_objects as go
          import plotly.offline as po
          from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot
          import plotly.express as px
          import random
          import plotly.figure_factory as ff
          from plotly.subplots import make_subplots
          from statsmodels.graphics.gofplots import qqplot
```



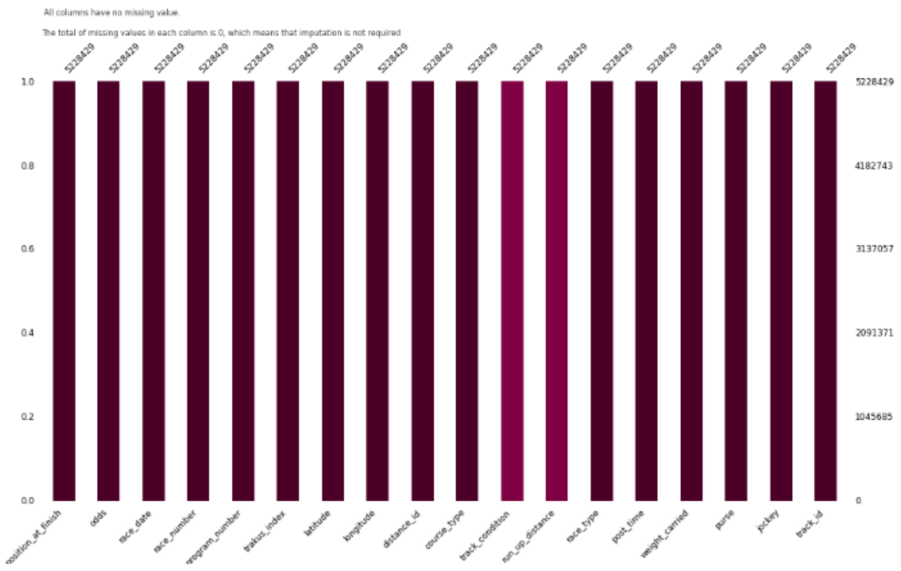
```
In [31]:  nyra_tracking_table = pd.read_csv("~/Desktop/Big Data Derby/big-data-derby-2022/nyra_tracking_table.csv")
          nyra_start_table=pd.read_csv("~/Desktop/Big Data Derby/big-data-derby-2022/nyra_start_table.csv")
          nyra_race_table=pd.read_csv("~/Desktop/Big Data Derby/big-data-derby-2022/nyra_race_table.csv")
          nyra_2019_complete=pd.read_csv("~/Desktop/Big Data Derby/big-data-derby-2022/nyra_2019_complete.csv", header = 0)
```

Data Preprocessing

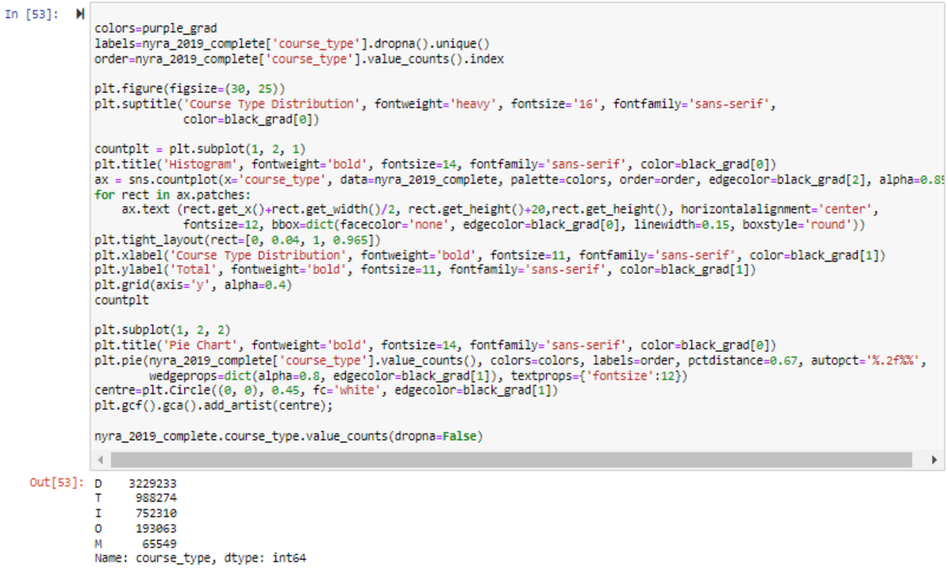


```
mso.bar(nyra_2019_complete, fontsize=9, color=[purple_grad[0], purple_grad[0], purple_grad[0], purple_grad[0], purple_grad[0]
                                   purple_grad[0], purple_grad[0], purple_grad[0], purple_grad[0], purple_grad[1], purple_grad[1]
        figsize=(15, 8), sort='descending', labels=True)

plt.suptitle('Missing Values in each Columns', fontweight='heavy', x=0.124, y=1.22, ha='left',fontsize='16',
             fontfamily='sans-serif', color=black_grad[0])
plt.title(' All columns have no missing value.\n\nThe total of missing values in each column is 0, which means that imputatic
          fontsize='8', fontfamily='sans-serif', loc='left', color=black_grad[1], pad=5)
plt.grid(axis='both', alpha=0);

nyra_2019_complete.isnull().sum()
```

```
track_id             0
race_date            0
race_number          0
program_number       0
trakus_index         0
latitude             0
longitude            0
distance_id          0
course_type          0
track_condition      0
run_up_distance      0
race_type            0
post_time            0
weight_carried       0
purse                0
jockey               0
odds                 0
position_at_finish   0
dtype: int64
```



Exploratory Data Analysis

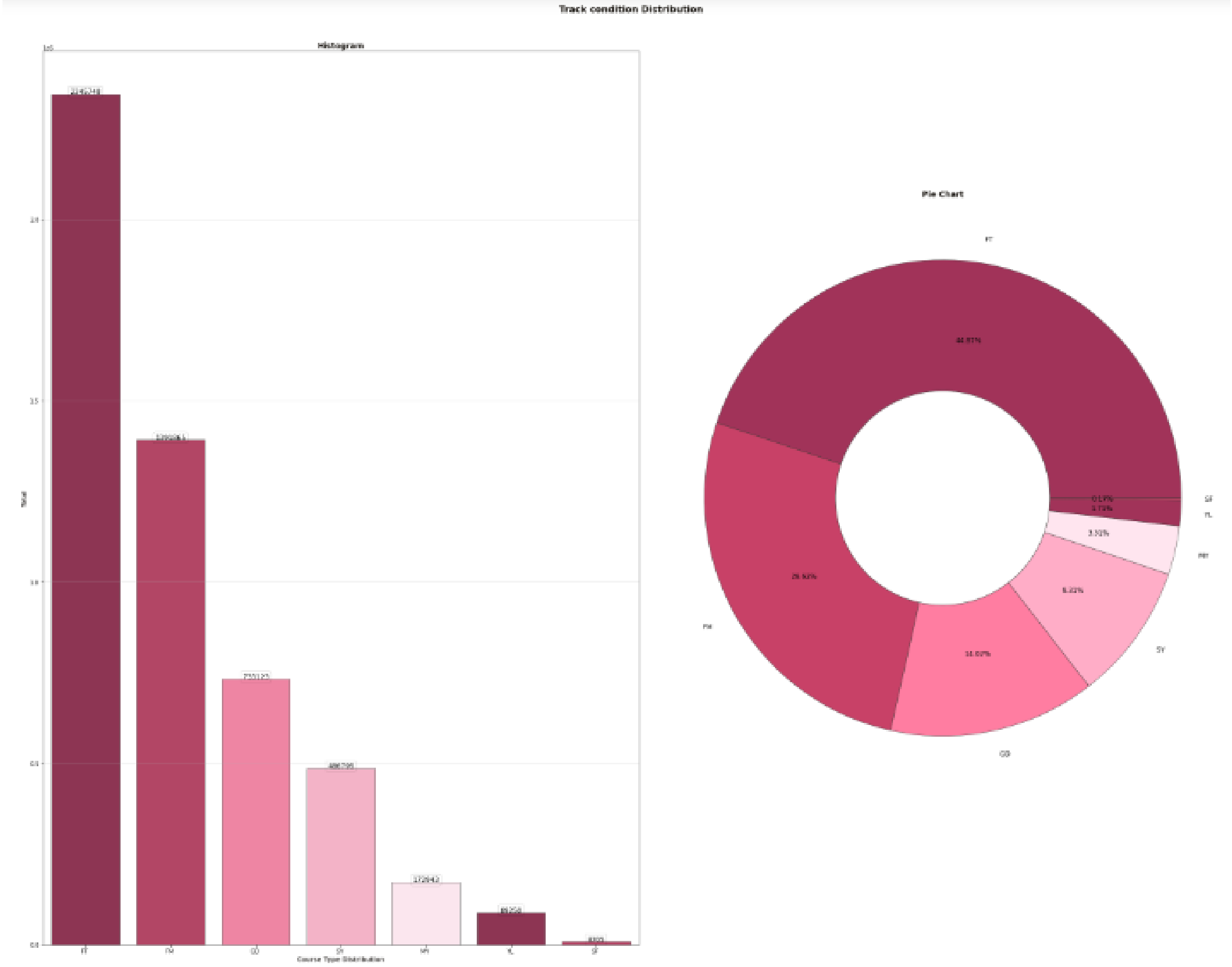


In [53]:

```
colors=purple_grad
labels=nyra_2019_complete['course_type'].dropna().unique()
order=nyra_2019_complete['course_type'].value_counts().index

plt.figure(figsize=(30, 25))
plt.suptitle('Course Type Distribution', fontweight='heavy', fontsize='16', fontfamily='sans-serif',
             color=black_grad[0])

countplt = plt.subplot(1, 2, 1)
plt.title('Histogram', fontweight='bold', fontsize=14, fontfamily='sans-serif', color=black_grad[0])
ax = sns.countplot(x='course_type', data=nyra_2019_complete, palette=colors, order=order, edgecolor=black_grad[2], alpha=0.85
for rect in ax.patches:
    ax.text (rect.get_x()+rect.get_width()/2, rect.get_height()+20,rect.get_height(), horizontalalignment='center',
             fontsize=12, bbox=dict(facecolor='none', edgecolor=black_grad[0], linewidth=0.15, boxstyle='round'))
plt.tight_layout(rect=[0, 0.04, 1, 0.965])
plt.xlabel('Course Type Distribution', fontweight='bold', fontsize=11, fontfamily='sans-serif', color=black_grad[1])
plt.ylabel('Total', fontweight='bold', fontsize=11, fontfamily='sans-serif', color=black_grad[1])
plt.grid(axis='y', alpha=0.4)
countplt

plt.subplot(1, 2, 2)
plt.title('Pie Chart', fontweight='bold', fontsize=14, fontfamily='sans-serif', color=black_grad[0])
plt.pie(nyra_2019_complete['course_type'].value_counts(), colors=colors, labels=order, pctdistance=0.67, autopct='%.2f%%',
        wedgeprops=dict(alpha=0.8, edgecolor=black_grad[1]), textprops={'fontsize':12})
centre=plt.Circle((0, 0), 0.45, fc='white', edgecolor=black_grad[1])
plt.gcf().gca().add_artist(centre);

nyra_2019_complete.course_type.value_counts(dropna=False)
```

Out[53]:

```
D    3229233
T     988274
I     752310
O     193063
M      65549
Name: course_type, dtype: int64
```



References:

https://www.researchgate.net/publication/50946368_Exploratory_data_analysis_in_the_context_of_data_mining_and_resampling

https://www.researchgate.net/publication/308007227_Exploratory_Data_Analysis
https://www.researchgate.net/publication/359182169_Optimal_Model_of_Horse_Racing_Competition_Decision_Management_Based_on_Association_Rules_and_Neural_Network