Summary

Nowadays data analysis has become increasingly popular and the conclusion from this process is strongly required in many fields, including Internet applications. In this paper, we proposed a model to develop a rating and ranking system.

First, we operated on the data provided. We finished data screening and data imputation in order to settle down to problem caused by missing value. After that, by normalization, we eliminated the difference between dimensions. After data processing, we obtained the data prepared for following analysis.

Next an algorithm was created to provide a rating/ranking criterion. Our method was based on TOPSIS method which is a valid approach for multi-index evaluation and decision making.

An additional algorithm was implied in our model to make a judgement about the quality of different roller coasters in a new way. This method is adapted from Rank Sun Ration method (RSR). RSR nowadays is being widely used in fields of evaluation and prediction.

Based on the two basic algorithms mentioned above, we built our model to

**Contents**

[1 Introduction 3](#_Toc529389692)

[1.1 Background 3](#_Toc529389693)

[1.2 Overview of our Work 3](#_Toc529389694)

[1.3 Assumptions 3](#_Toc529389695)

[2 Data Processing 4](#_Toc529389696)

[2.1 Data Screening 4](#_Toc529389697)

[2.2 Data Imputation 4](#_Toc529389698)

[2.3 Data Normalization 4](#_Toc529389699)

[2.4 Data Whitening 4](#_Toc529389700)

[2.5 Data Visualization 4](#_Toc529389701)

[3 ANN Evalutation System 5](#_Toc529389702)

[3.1 Concept of PCA 5](#_Toc529389703)

[3.2 Concept of ANN 5](#_Toc529389704)

[4 Model Construction 6](#_Toc529389705)

[4.1 Definition of Risk 6](#_Toc529389706)

[4.2 Basic Model 6](#_Toc529389707)

[4.3 Result of basic model 6](#_Toc529389708)

[4.4 Error analysis 6](#_Toc529389709)

[5 Sensitive analysis and validation 7](#_Toc529389710)

[5.1 Risk Return 7](#_Toc529389711)

[5.2 Attribute 7](#_Toc529389712)

[5.3 Policy of distribution 7](#_Toc529389713)

[6 Future Work 8](#_Toc529389714)

[7 Conclusion 9](#_Toc529389715)

[7.1 Strengths 9](#_Toc529389716)

[7.2 Weaknesses 9](#_Toc529389717)

# 1 Introduction

## 1.1 Background

With the rapid development of the society, people nowadays are enjoying a higher living condition. It is wildly known that going to the theme park is always a wise option during the weekend or the holiday. Among those facilities the roller coaster is always the most exciting one. What is more, many people even would like to choose a certain theme park simply because the roller coaster there. As a result, the different features of roller coaters are always utilized in the design for the leaflet of the amusement park for advertisements.

The twenty first century is an epoch of the Internet. With the convenient access to the Internet, huge amount of data has become available for everyone with merely several clicks. In other words, we are able to obtain the information of theme parks as well as roller coasters all over the world. Moreover, the unpresented economic cooperation between different countries has made traveling abroad more and more appealing. In the past there are not so many foreign tourists in the amusement park at all. So there is a strong demand for a new general rating / ranking system which can provide every individual accurate and considerate recommendation on roller coasters all over the world.

The article is about a set of algorithms developing a descriptive roller coaster rating/ranking system.

## 1.2 Overview of our Work

## 1.3 Assumptions

# 2 Data Processing

Any serious statistical analysis begins with substantial work on data. In this section, we adapted several measures to do with random missingness and unnatural variables.

## 2.1 Data Screening

First, we do data screening on 300 roller coasters:

* After basic statistic, we find that there is large amounts of missingness in four columns: Drop(feet), G Force , Duration(min : sec) and Vertical Angle(degrees). The “Drop” column has 159 empty value in total, the “G Force” has 217 missing data , as for “Duration” there are 77 missing data and there is 209 absence found in “Vertical Angle”.
* Only slight absence are found in Height, Speed and Length. Apart from columns mentioned, other columns have complete information.

Second, we use Pearson Correlation to reflect the relationship between different attributions:

* MATLAB has many useful tools for dealing with missing data and mapping. Following figures show the result of using Pearson Correlation to analyze three columns: Drop, G force, Duration, and Vertical Angle. In these figures, x axis represents properties (columns) and y axis represents Pearson Correlation Coefficient.
* In Figure 1, which reflects the relationship between Drop and other eight columns, indicates that Drop is closely connected with Length and Height. However, the other three figures( Figure 1, Figure 2, Figure 3) suggest that each of these three attributes (G Force, Duration and Vertical Angle) has inapparent connection with others. That means, their position can not be replaced by others. In other words, it is not reasonable to delete any single attribute.

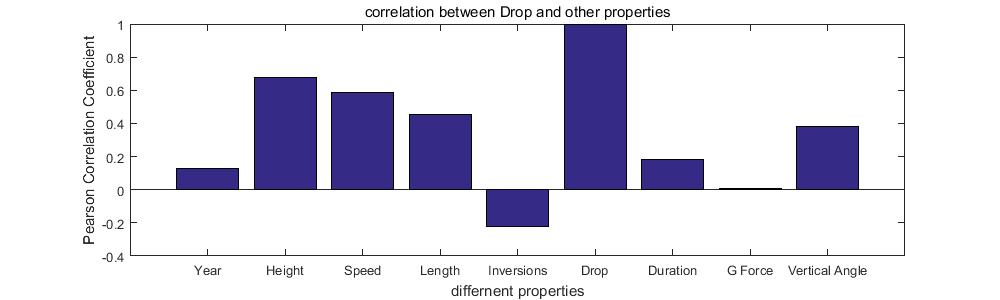


Figure 1 : Pearson Correlation between Drop and other properties

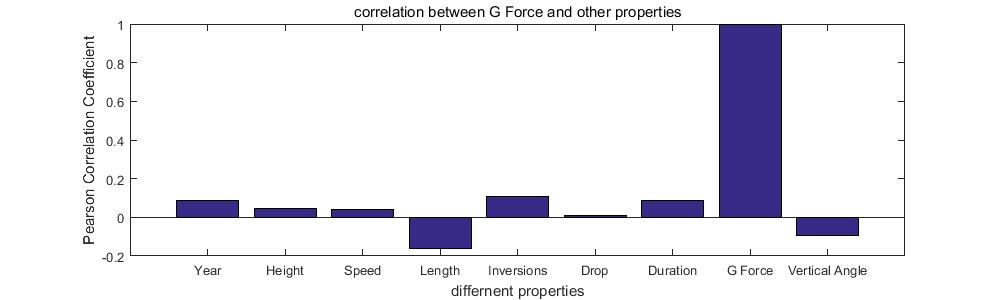


Figure 2 : Pearson Correlation between G Force and other properties

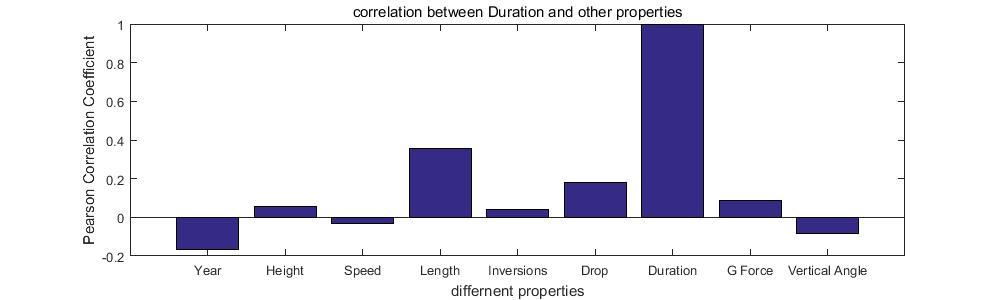


Figure 3 : Pearson Correlation between Duration and other properties

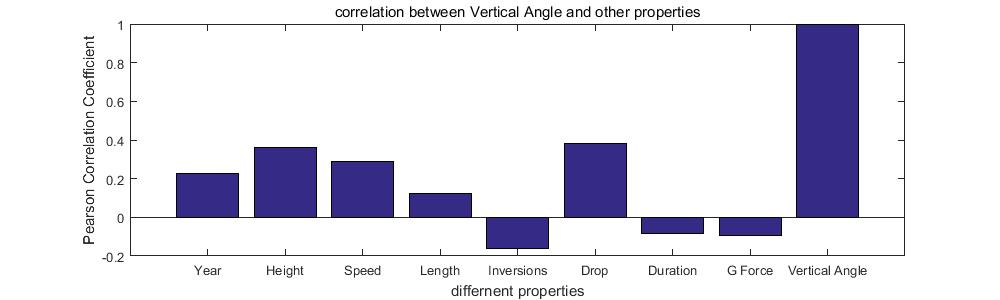


Figure 4: Pearson Correlation between Vertical Angle and other properties

## 2.2 Data Imputation

## 2.3 Data Normalization

## 2.4 Data Whitening

## 2.5 Data Visualization

# 3 ANN Evalutation System

## 3.1 Concept of PCA

## 3.2 Concept of ANN

# 4 Model Construction

## 4.1 Definition of Risk

## 4.2 Basic Model

## 4.3 Result of basic model

## 4.4 Error analysis

# 5 Sensitive analysis and validation

## 5.1 Risk Return

## 5.2 Attribute

## 5.3 Policy of distribution

# 6 Future Work

# 7 Conclusion

## 7.1 Strengths

## 7.2 Weaknesses