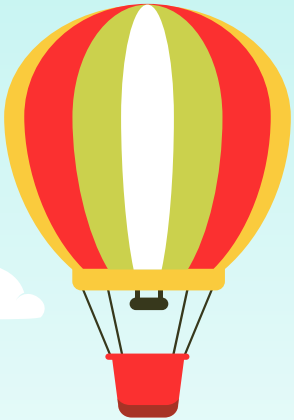
Rollercoaster Dataset EDA PPT

TABLE OF CONTENTS

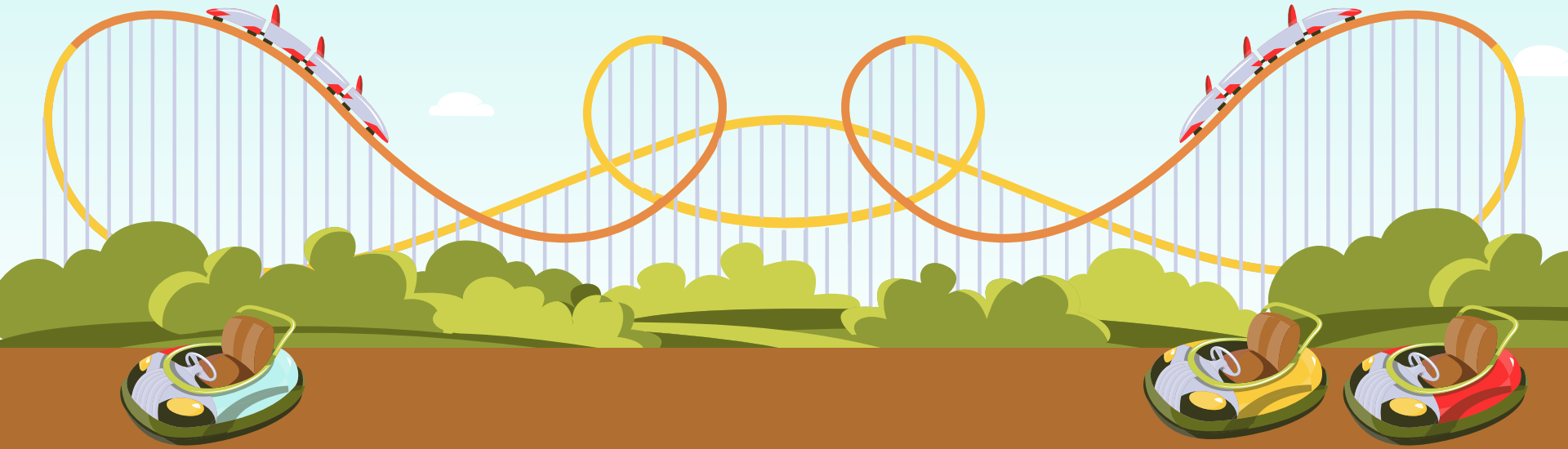


01 OVERVIEW

Here We'll Take a look at the dataset in a brief to understand why this dataset is used and why have we conducted this EDA.

Dataset Overview

The dataset consists of 1,087 records of roller coasters, each described by 56 variables. These variables include both categorical attributes (e.g., Location, Manufacturer, Status) and numerical attributes (e.g., Height, Speed, Year Introduced). This variety in data types necessitates a careful approach to data cleaning and preprocessing to ensure the accuracy and reliability of subsequent analyses.



Objective

The primary objective of this analysis is to perform an in-depth exploration of a roller coaster dataset. This dataset encompasses various attributes of roller coasters, including their locations, operating statuses, technical specifications, and manufacturers.

By analyzing this data, we aim to identify trends in roller coaster designs, understand the factors that influence their performance, and uncover insights that could be useful for theme park operators, ride designers, and enthusiasts.

KEY QUESTIONS ANSWERED

1. What is the relationship between speed and height in roller coasters?
2. How has roller coaster design evolved over the years?
3. What are the most common speed ranges, and how are they distributed?
4. Which years saw the highest number of roller coaster introductions?

And more....

02

DATA CLEANING

Now that we have an idea about the dataset, Let's look at how it has been processed to get the best results.

Data Cleaning and Preparation

1. Initial Data Inspection
2. Handling Missing Values
3. Data Type Conversion
4. Outlier Detection and Treatment

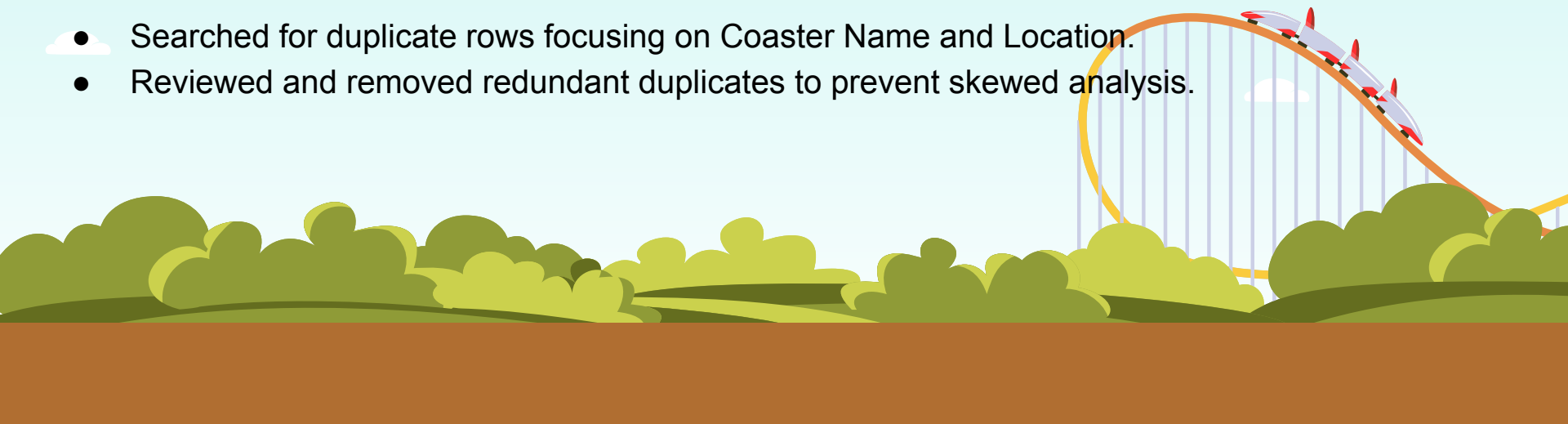
Initial Data Inspection

Loading and Previewing the Data:

- Loaded dataset into pandas DataFrame.
- Previewed the first few rows to understand data structure.
- Detected initial data quality issues like missing values and inconsistent data types.

Checking for Duplicates:

- Searched for duplicate rows focusing on Coaster Name and Location.
- Reviewed and removed redundant duplicates to prevent skewed analysis.



Handling Missing Values

Identification of Missing Values:

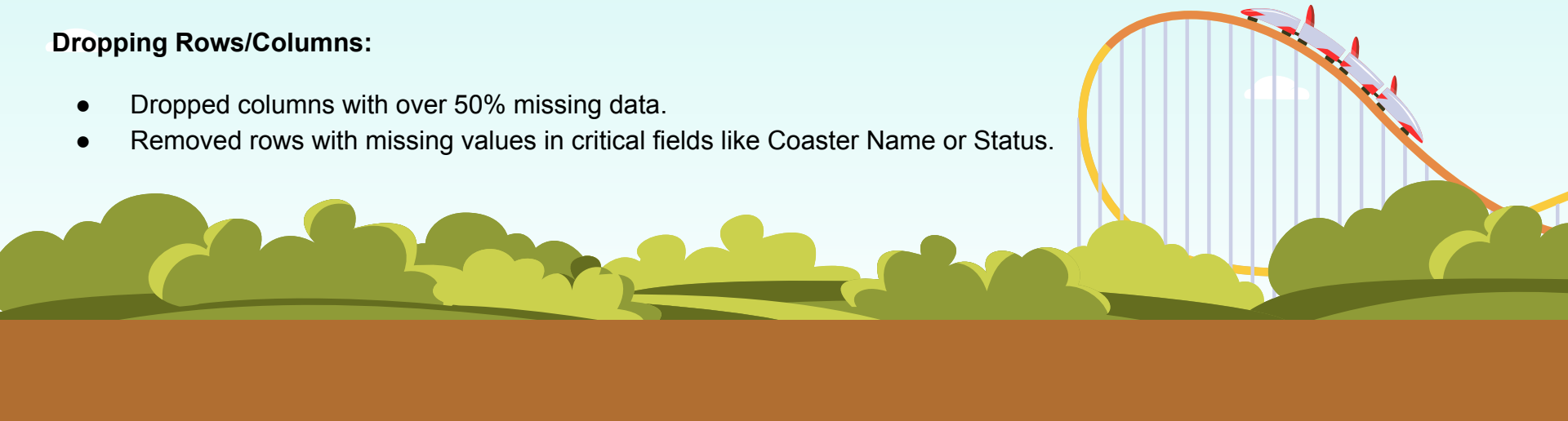
- Used `df.info()` and `df.isnull().sum()` to quantify missing data.
- Noticed missing values in critical columns like Height, Speed, and Inversions.

Strategy for Imputing Missing Data:

- **Numerical Columns:** Imputed missing values using the median to minimize the effect of outliers.
- **Categorical Columns:** Filled missing values with placeholders ("Unknown") or mode.

Dropping Rows/Columns:

- Dropped columns with over 50% missing data.
- Removed rows with missing values in critical fields like Coaster Name or Status.



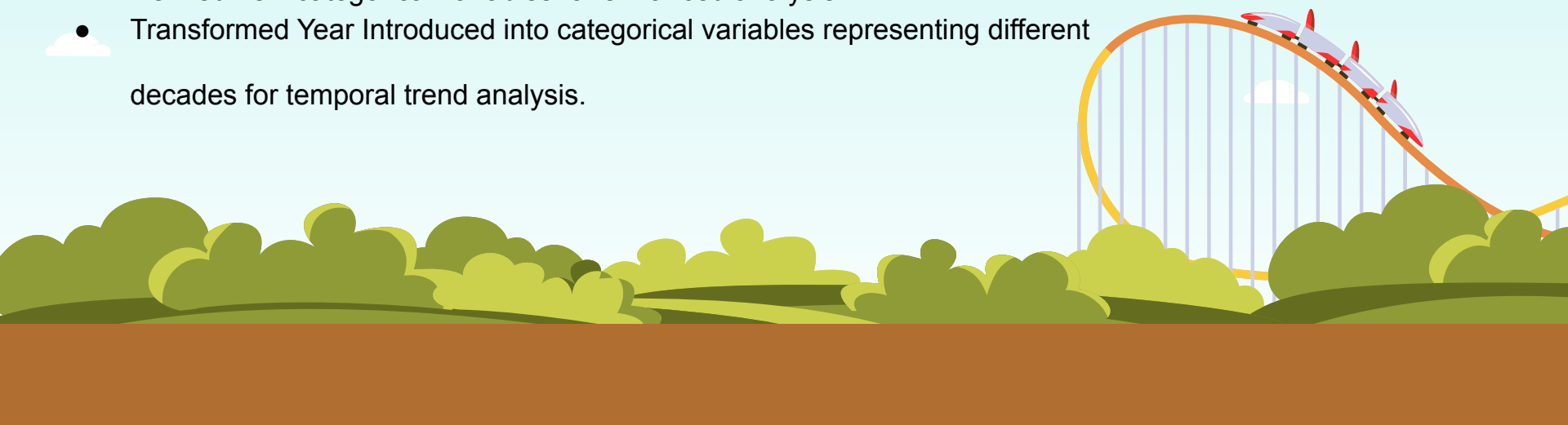
Data Type Conversion

Correcting Data Types:

- Identified incorrect data types during initial inspection.
- Used `pd.to_numeric()` and `pd.to_datetime()` to convert columns like Year Introduced, Height, and Speed to appropriate types.

Creation of New Categorical Variables:

- Derived new categorical variables for enhanced analysis.
- Transformed Year Introduced into categorical variables representing different decades for temporal trend analysis.



Outlier Detection and Treatment

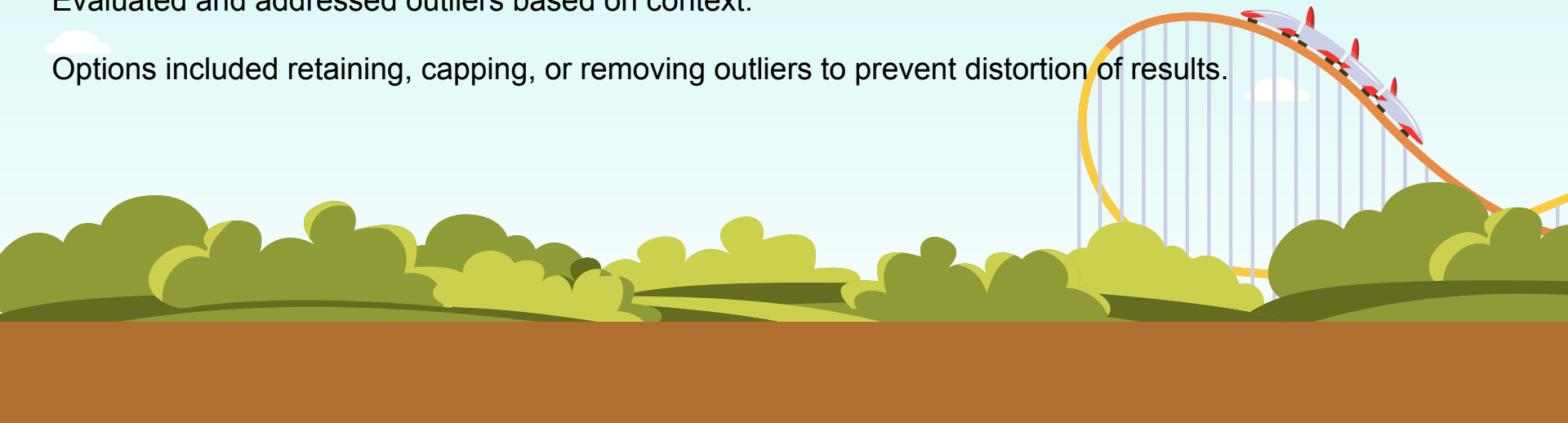
Identifying Outliers:

Used box plots and statistical methods like Z-score to detect outliers in Height, Speed, and Inversions.

Handling Outliers:

Evaluated and addressed outliers based on context.

Options included retaining, capping, or removing outliers to prevent distortion of results.



03

VISUALISATIONS

Top Years for Roller Coaster Introductions

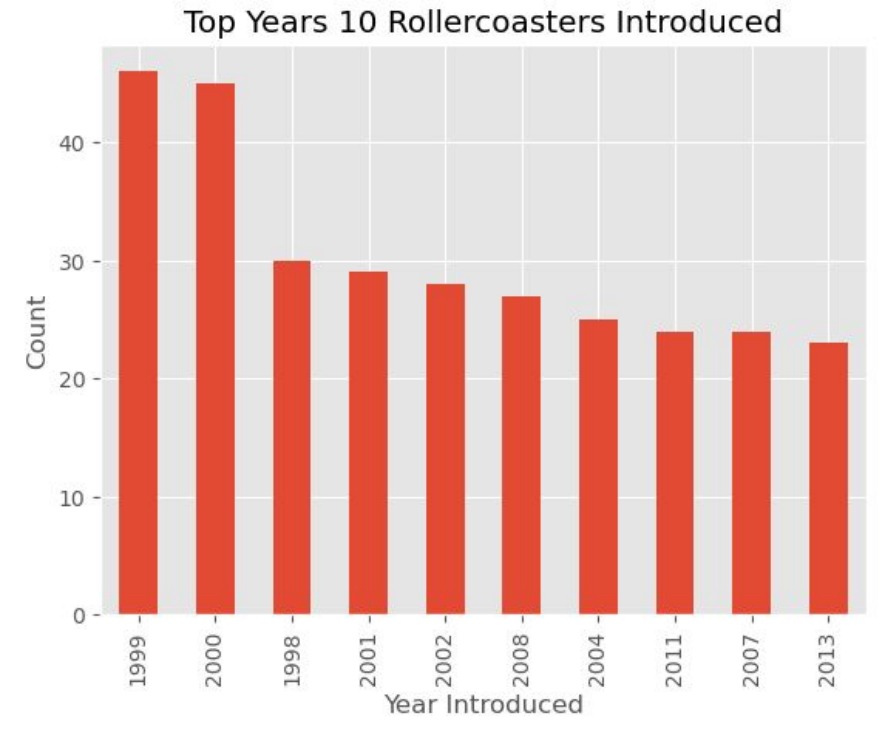
Presenting: Timeline of Roller Coaster Introductions

Question Being Answered: What were the most active years for new roller coaster launches?

Key Insights:

- The year 2001 saw the introduction of 30 new roller coasters.
- Significant activity was also observed in 1999 and 2000, with nearly 45-44 introductions..
- The data indicates key growth periods in the roller coaster industry.

Importance: Understanding the timeline of roller coaster introductions provides valuable context for industry growth trends and helps predict future innovations in roller coaster design.



Histogram - Roller Coaster Speed Distribution

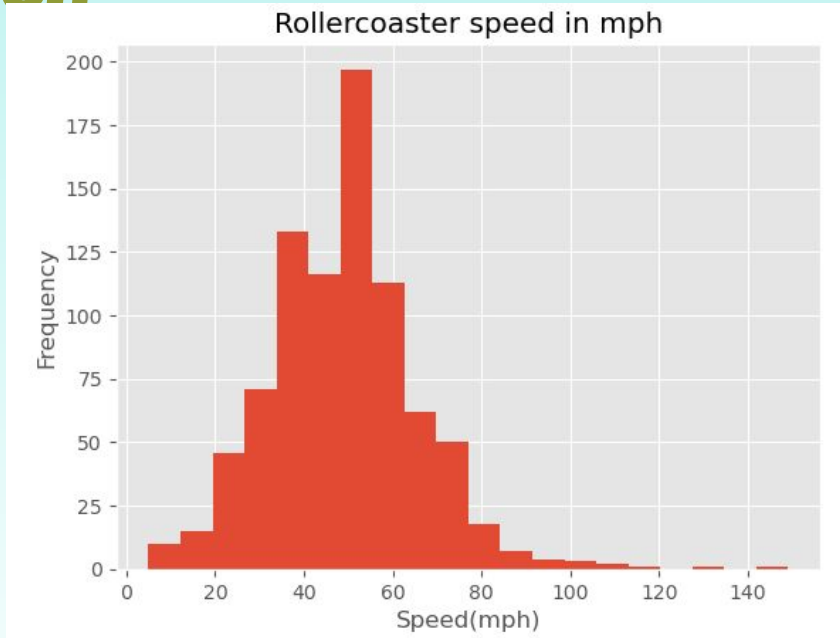
Presenting: Distribution of Roller Coaster Speeds

Question Being Answered: What is the distribution of roller coaster speeds? Are there common speed ranges, and how do they vary?

Key Insights:

- The histogram shows that most roller coasters have speeds between **50 and 100 mph**, with a peak around **70 mph**.
- A few outliers exceed **120 mph**.

Importance: Understanding the distribution of speeds is critical for recognizing industry standards and how they might evolve. This information is crucial for designing future roller coasters and setting benchmarks for safety and thrill.



Scatter Plot - Speed vs. Height

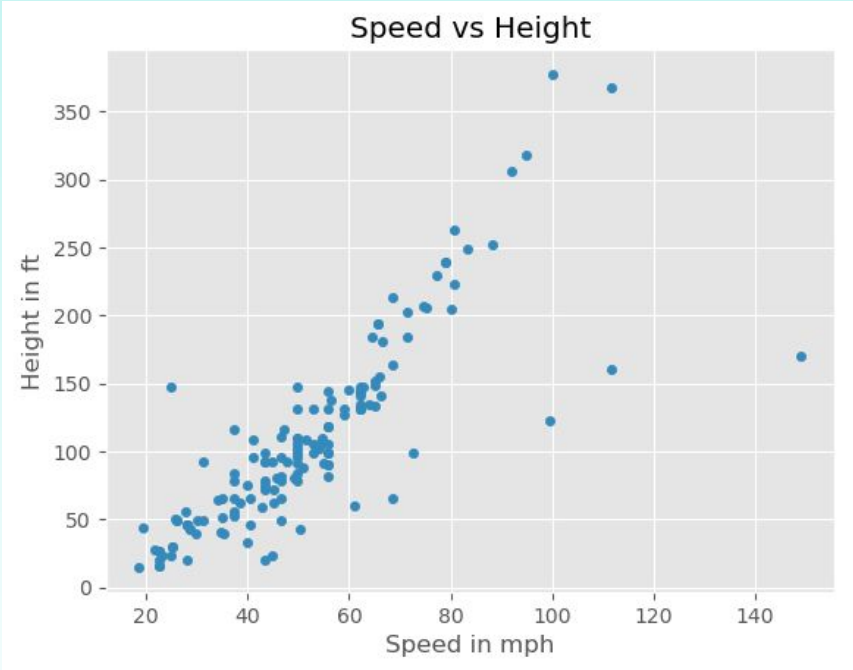
Presenting: Relationship Between Speed and Height in Roller Coasters

Question Being Answered: What is the relationship between the speed and height of roller coasters? Are taller roller coasters generally faster?

Key Insights:

- The scatter plot shows a positive correlation between speed and height.
- Coasters with speeds around **100 mph** tend to have heights close to **300 ft**.
- Taller roller coasters are often designed to achieve higher speeds.

Importance: This relationship is crucial in understanding the design trends in roller coasters. Highlighting this trend shows how designers leverage height to increase speed, a key factor in creating thrilling rides. Understanding this helps in predicting future designs and assessing safety measures.



Scatter Plot with Color-coded Year Introduced

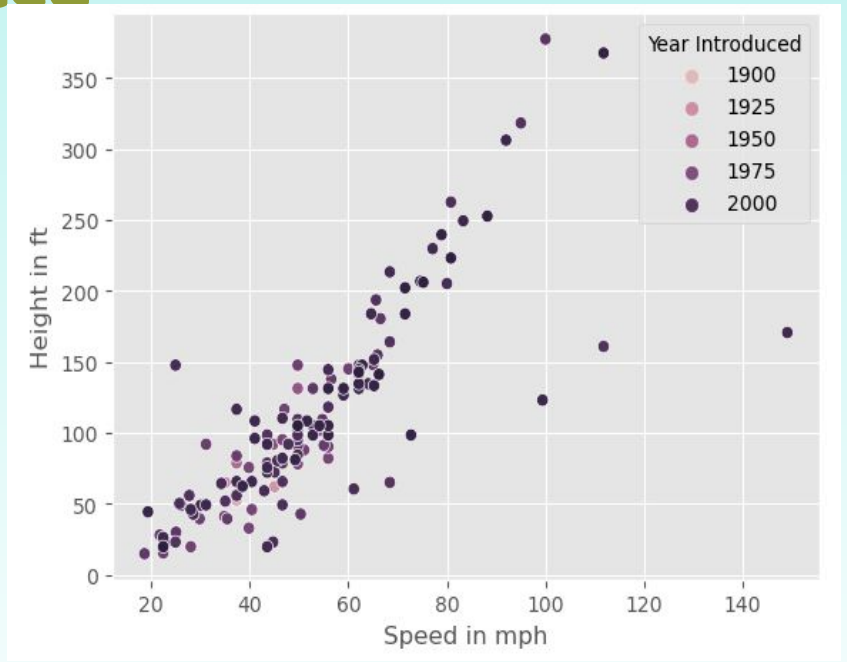
Presenting: Evolution of Roller Coaster Design Over Time

Question Being Answered: How does the relationship between speed and height vary across roller coasters introduced in different years? Are newer coasters faster and taller?

Key Insights:

- The scatter plot color-coded by year indicates that newer coasters, particularly those introduced in later 2000s, are generally faster and taller.
- Coasters introduced later in time are often positioned at the higher end of both speed and height.

Importance: This trend illustrates the technological advancements in roller coaster design, driven by the demand for more thrilling experiences. Understanding this progression helps in forecasting future industry trends and aligning investment strategies with these advancements.



Pair Plot - Interaction Between Multiple Variables

Presenting: Interaction Between Speed, Height, and Year Introduced

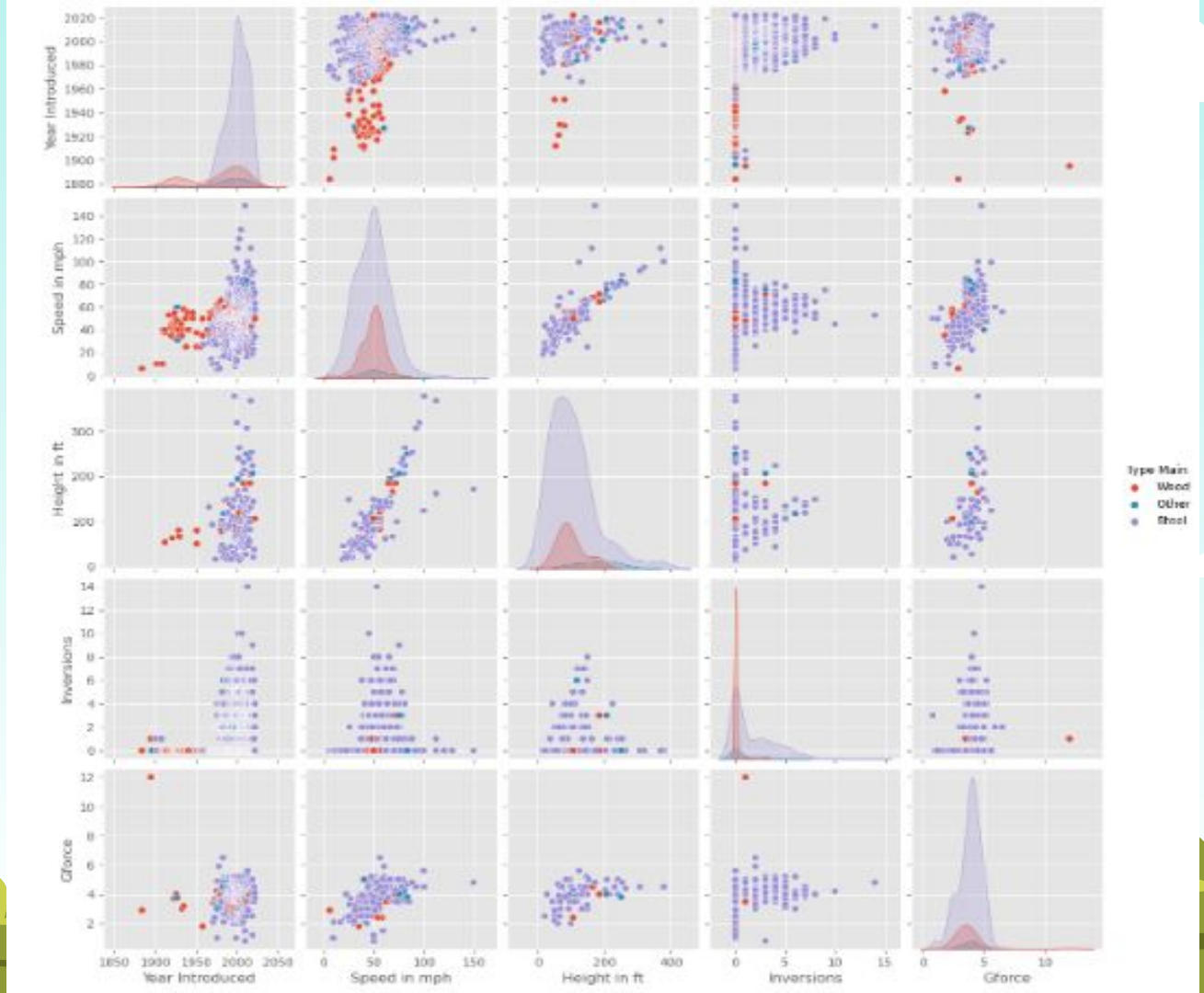
Question Being Answered: How do variables like speed, height, and year introduced interact with each other? Are there any strong correlations or patterns?

Key Insights:

- The pair plot reveals strong correlations, especially between speed and height.
- The plot shows that roller coasters have generally become taller and faster over time, with each variable influencing the others.

Importance: A comprehensive understanding of these interactions is essential for identifying key trends in roller coaster design. This multivariable analysis helps in pinpointing the factors that most significantly influence roller coaster characteristics.

Pair Plot – Interaction Between Multiple Variables



Correlation Heatmap

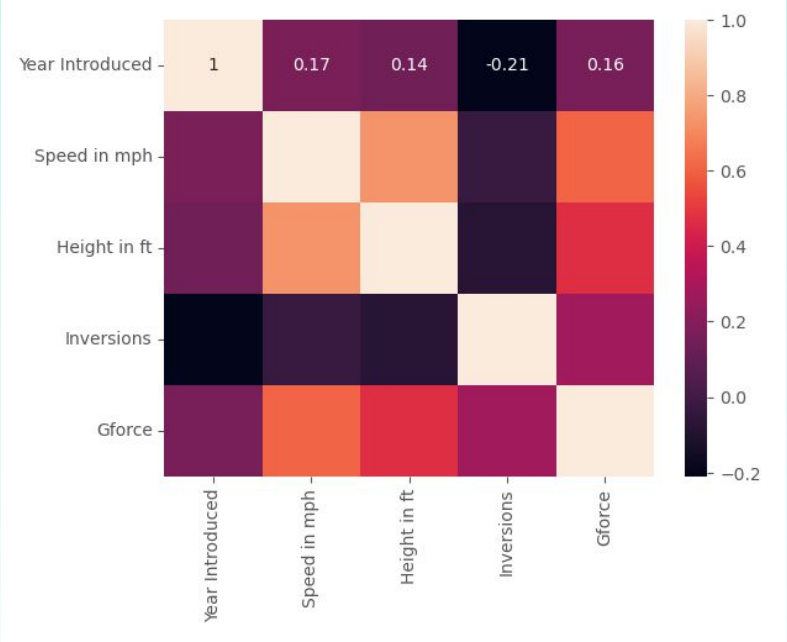
Presenting: Relationship Between Speed and Height in Roller Coasters

Question Being Answered: What is the relationship between the speed and height of roller coasters? Are taller roller coasters generally faster?

Key Insights:

- The heatmap shows a positive correlation between speed and height.
- Taller roller coasters tend to achieve higher speeds.
- The relationship between height and inversions is weaker.

Importance: Understanding the correlation between speed and height is crucial in roller coaster design, as it shows how these two factors work together to create thrilling rides. This insight can help predict future design trends and assess safety measures.



Avg Roller Coaster Speed by Location

Presenting: Distribution of roller coaster speed based on their location

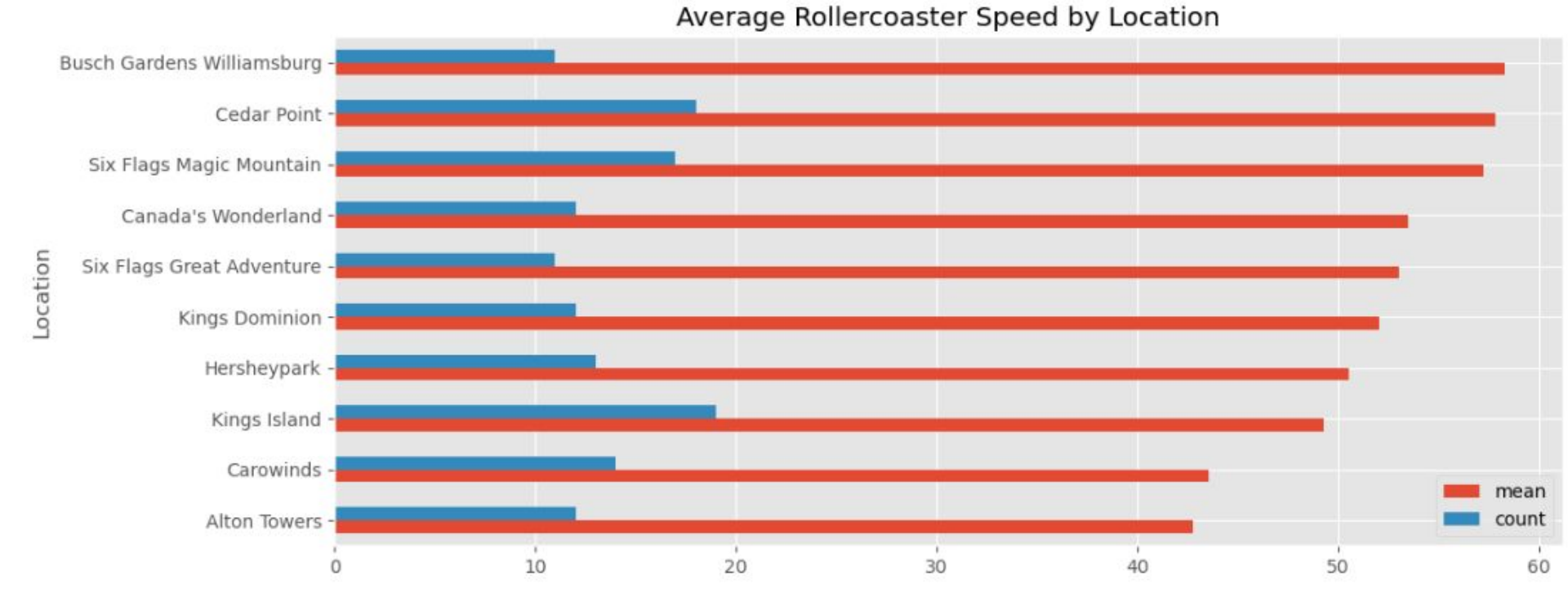
Question Being Answered: Which locations boast the highest roller coaster speed?

Key Insights:

- Busch Gardens Williamsburg, Cedar Point and Six Flags Magic Mountain are leading.
- The distribution highlights major amusement park hubs globally.

Importance: Identifying where fastest and strongest roller coasters are located provides insight into regional market saturation and highlights key locations that attract significant visitor numbers.

Avg Roller Coaster Speed by



CONCLUSION

04

CONCLUSION

The analysis of the roller coaster dataset reveals several important insights that enhance our understanding of roller coaster design trends and industry growth. Through various visualizations, including scatter plots, pair plots, histograms, and bar charts, we identified key relationships and trends in the data.

1. **Speed vs. Height Correlation:** There is a clear positive correlation between the speed and height of roller coasters. This relationship indicates that taller roller coasters are often designed to achieve higher speeds, leveraging height to enhance the thrill factor. This trend is significant for understanding how roller coaster designs have evolved to meet the demand for more exhilarating experiences.
2. **Evolution Over Time:** By incorporating the year of introduction into the analysis, it becomes evident that roller coasters have generally become faster and taller over time. This progression reflects technological advancements and a growing emphasis on creating more intense rides. The trend also highlights how the industry has responded to consumer demand for heightened excitement.
3. **Multiple Variable Interaction:** The pair plot analysis confirmed that multiple variables, such as speed, height, and year introduced, are interrelated. These relationships underscore the complex considerations that go into roller coaster design, where factors like speed and height are carefully balanced to create optimal experiences.

CONCLUSION

1. **Speed Distribution:** The distribution of roller coaster speeds reveals that most coasters fall within the 50-100 mph range, with a peak around 70 mph. This concentration suggests a standard or preferred speed range, likely balancing the thrill of higher speeds with safety considerations.
2. **Geographical Distribution:** The analysis of roller coaster speed by location highlights the prominence of certain amusement parks, such as Busch Gardens Williamsburg, Cedar Point and Six Flags Magic Mountain, as major hubs for roller coasters. This information is valuable for understanding regional trends and the competitive landscape in the amusement park industry.
3. **Peak Years of Introduction:** The analysis of roller coaster introductions over time shows that the early 2000s were peak years for new roller coasters. This period likely represents a convergence of technological innovation and increased consumer demand, leading to a surge in new coaster designs.

In summary, this exploratory data analysis provides a comprehensive view of the roller coaster industry, illustrating key trends in design, distribution, and innovation. These insights are crucial for industry stakeholders, including designers, engineers, and amusement park operators, as they navigate future developments and continue to push the boundaries of what roller coasters can offer.