Data in Motion Pandas Challenge Week 7

import

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Read Data

In [2]:

```
url = 'https://raw.githubusercontent.com/kedeisha1/Challenges/main/coaster_db.csv'

df = pd.read_csv(url)
```

In [3]:

```
df.head()
```

Out[3]:

	coaster_name	Length	Speed	Location	Status	Opening date	Type	Manufacturer	Height restriction	Model
0	Switchback Railway	600 ft (180 m)	6 mph (9.7 km/h)	Coney Island	Removed	June 16, 1884	Wood	LaMarcus Adna Thompson	NaN	Lift Packed
1	Flip Flap Railway	NaN	NaN	Sea Lion Park	Removed	1895	Wood	Lina Beecher	NaN	NaN
2	Switchback Railway (Euclid Beach Park)	NaN	NaN	Cleveland, Ohio, United States	Closed	NaN	Other	NaN	NaN	NaN
3	Loop the Loop (Coney Island)	NaN	NaN	Other	Removed	1901	Steel	Edwin Prescott	NaN	NaN
4	Loop the Loop (Young's Pier)	NaN	NaN	Other	Removed	1901	Steel	Edwin Prescott	NaN	NaN

5 rows × 56 columns

In [4]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1087 entries, 0 to 1086
Data columns (total 56 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   coaster_name                    1087 non-null   object
 1   Length                          953 non-null    object
 2   Speed                           937 non-null    object
 3   Location                        1087 non-null   object
 4   Status                          874 non-null    object
 5   Opening date                    837 non-null    object
 6   Type                            1087 non-null   object
 7   Manufacturer                    1028 non-null   object
 8   Height restriction              831 non-null    object
 9   Model                           744 non-null    object
 10  Height                          965 non-null    object
 11  Inversions                      932 non-null    float64
 12  Lift/launch system              795 non-null    object
 13  Cost                            382 non-null    object 
 14  Trains                          718 non-null    object 
 15  Park section                    487 non-null    object 
 16  Duration                        765 non-null    object 
 17  Capacity                        575 non-null    object 
 18  G-force                         362 non-null    object 
 19  Designer                        578 non-null    object 
 20  Max vertical angle              357 non-null    object 
 21  Drop                            494 non-null    object 
 22  Soft opening date               96 non-null     object 
 23  Fast Lane available             69 non-null     object 
 24  Replaced                        173 non-null    object 
 25  Track layout                    335 non-null    object 
 26  Fastrack available              19 non-null     object 
 27  Soft opening date.1             96 non-null     object 
 28  Closing date                    236 non-null    object 
 29  Opened                          27 non-null     object 
 30  Replaced by                     88 non-null     object 
 31  Website                         87 non-null     object 
 32  Flash Pass Available            50 non-null     object 
 33  Must transfer from wheelchair   106 non-null    object 
 34  Theme                           44 non-null     object 
 35  Single rider line available     81 non-null     object 
 36  Restraint Style                 22 non-null     object 
 37  Flash Pass available            46 non-null     object 
 38  Acceleration                    60 non-null     object 
 39  Restraints                      24 non-null     object 
 40  Name                            35 non-null     object 
 41  year_introduced                 1087 non-null   int64  
 42  latitude                        812 non-null    float64
 43  longitude                       812 non-null    float64
 44  Type_Main                       1087 non-null   object 
 45  opening_date_clean              837 non-null    object 
 46  speed1                          937 non-null    object 
 47  speed2                          935 non-null    object 
 48  speed1_value                    937 non-null    float64
 49  speed1_unit                     937 non-null    object 
 50  speed_mph                       937 non-null    float64
 51  height_value                    965 non-null    float64
 52  height_unit                     965 non-null    object 
 53  height_ft                       171 non-null    float64
 54  Inversions_clean                1087 non-null   int64  
 55  Gforce_clean                    362 non-null    float64
dtypes: float64(8), int64(2), object(46)
memory usage: 475.7+ KB
```

Q1. How many columns and rows are in the dataset?

In [5]:

```
print ("rows number",df.shape[0])
```

```
rows number 1087
```

In [6]:

```
print ("columns number",df.shape[1])
```

```
columns number 56
```

Q2. Is there any missing data?

In [7]:

```
df.isnull().sum()
```

Out[7]:

```
coaster_name                       0
Length                           134
Speed                            150
Location                           0
Status                           213
```

```
Opening date                    250
Type                              0
Manufacturer                     59
Height restriction              256
Model                           343
Height                          122
Inversions                      155
Lift/launch system              292
Cost                            705
Trains                          369
Park section                    600
Duration                        322
Capacity                        512
G-force                         725
Designer                        509
Max vertical angle              730
Drop                            593
Soft opening date               991
Fast Lane available            1018
Replaced                        914
Track layout                    752
Fastrack available             1068
Soft opening date.1             991
Closing date                    851
Opened                         1060
Replaced by                     999
Website                        1000
Flash Pass Available           1037
Must transfer from wheelchair   981
Theme                          1043
Single rider line available    1006
Restraint Style                1065
Flash Pass available           1041
Acceleration                   1027
Restraints                     1063
Name                           1052
year_introduced                   0
latitude                        275
longitude                       275
Type_Main                         0
opening_date_clean              250
speed1                          150
speed2                          152
speed1_value                    150
speed1_unit                     150
speed_mph                       150
height_value                    122
height_unit                     122
height_ft                       916
Inversions_clean                  0
Gforce_clean                    725
dtype: int64
```

Q3. Display the summary statistics of the numeric columns using the describe method.

In [8]:

```
df.describe()
```

Out[8]:

	Inversions	year_introduced	latitude	longitude	speed1_value	speed_mph	height_value	height
count	932.000000	1087.000000	812.000000	812.000000	937.000000	937.000000	965.000000	171.0000
mean	1.547210	1994.986201	38.373484	-41.595373	53.850374	48.617289	89.575171	101.9964
std	2.114073	23.475248	15.516596	72.285227	23.385518	16.678031	136.246444	67.3290
min	0.000000	1884.000000	-48.261700	-123.035700	5.000000	5.000000	4.000000	13.1000

25%	0.000000	1989.000000	35.031050	-84.552200	40.000000	37.300000	44.000000	51.8000
50%	0.000000	2000.000000	40.289800	-76.653600	50.000000	49.700000	79.000000	91.2000
75%	3.000000	2010.000000	44.799600	2.778100	63.000000	58.000000	113.000000	131.2000
max	14.000000	2022.000000	63.230900	153.426500	240.000000	149.100000	3937.000000	377.3000

Q4. Rename the following columns:

- coaster_name ➡ Coaster_Name
- year_introduced ➡ Year_Introduced
- opening_date_clean ➡ Opening_Date
- speed_mph ➡ Speed_mph
- height_ft ➡ Height_ft
- Inversions_clean ➡ Inversions
- Gforce_clean ➡ Gforce

In [9]:

```
df.rename(columns={"coaster_name":"Coaster_Name","year_introduced":"Year_Introduced","op
```

In [10]:

```
df.columns
```

Out[10]:

```
Index(['Coaster_Name', 'Length', 'Speed', 'Location', 'Status', 'Opening date',
       'Type', 'Manufacturer', 'Height restriction', 'Model', 'Height',
       'Inversions', 'Lift/launch system', 'Cost', 'Trains', 'Park section',
       'Duration', 'Capacity', 'G-force', 'Designer', 'Max vertical angle',
       'Drop', 'Soft opening date', 'Fast Lane available', 'Replaced',
       'Track layout', 'Fastrack available', 'Soft opening date.1',
       'Closing date', 'Opened', 'Replaced by', 'Website',
       'Flash Pass Available', 'Must transfer from wheelchair', 'Theme',
       'Single rider line available', 'Restraint Style',
       'Flash Pass available', 'Acceleration', 'Restraints', 'Name',
       'Year_Introduced', 'latitude', 'longitude', 'Type_Main', 'Opening_Date',
       'speed1', 'speed2', 'speed1_value', 'speed1_unit', 'Speed_mph',
       'height_value', 'height_unit', 'Height_ft', 'Inversions', 'Gforce'],
      dtype='object')
```

Q5. Are there any duplicated rows?

In [11]:

```
sum(df.duplicated())
```

Out[11]:

```
0
```

Q6.What are the top 3 years with the most roller coasters introduced?

In [12]:

```
top3Years=np.array(df['Year_Introduced'].value_counts().sort_values(ascending=False).hea
top3Years
```

Out[12]:

```
array([1999, 2000, 1998])
```

Q7.What is the average speed? Also display a plot to show it's distribution.

In [13]:

```
print('speed average:',df.Speed_mph.mean(),'mph')
```

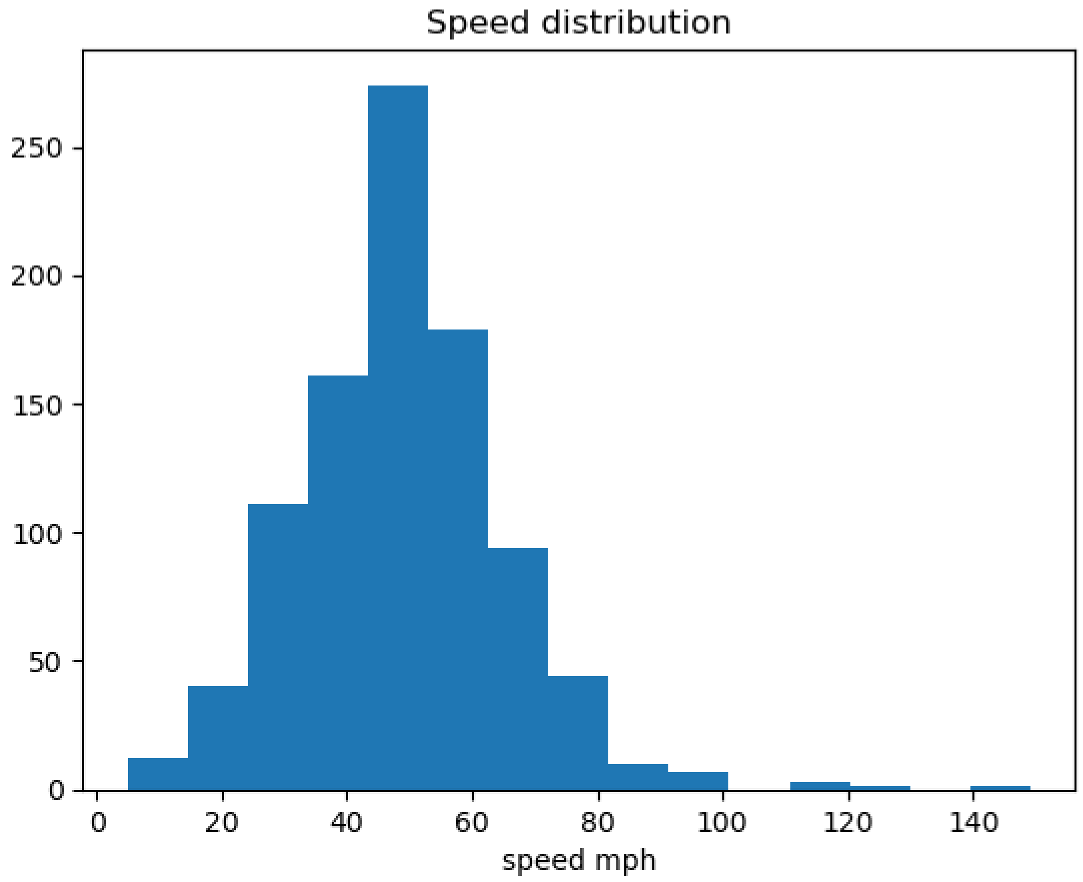
```
speed average: 48.617289220917804 mph
```

In [14]:

```
plt.hist(df.Speed_mph,bins=15)
plt.title("Speed distribution")
plt.xlabel('speed mph')
plt.show
```

Out[14]:

```
<function matplotlib.pyplot.show(close=None, block=None)>
```



Q8. Explore the feature relationships. Are there any positively or negatively correlated relationships?

In [15]:

```
correlation=df.corr()
correlation
```

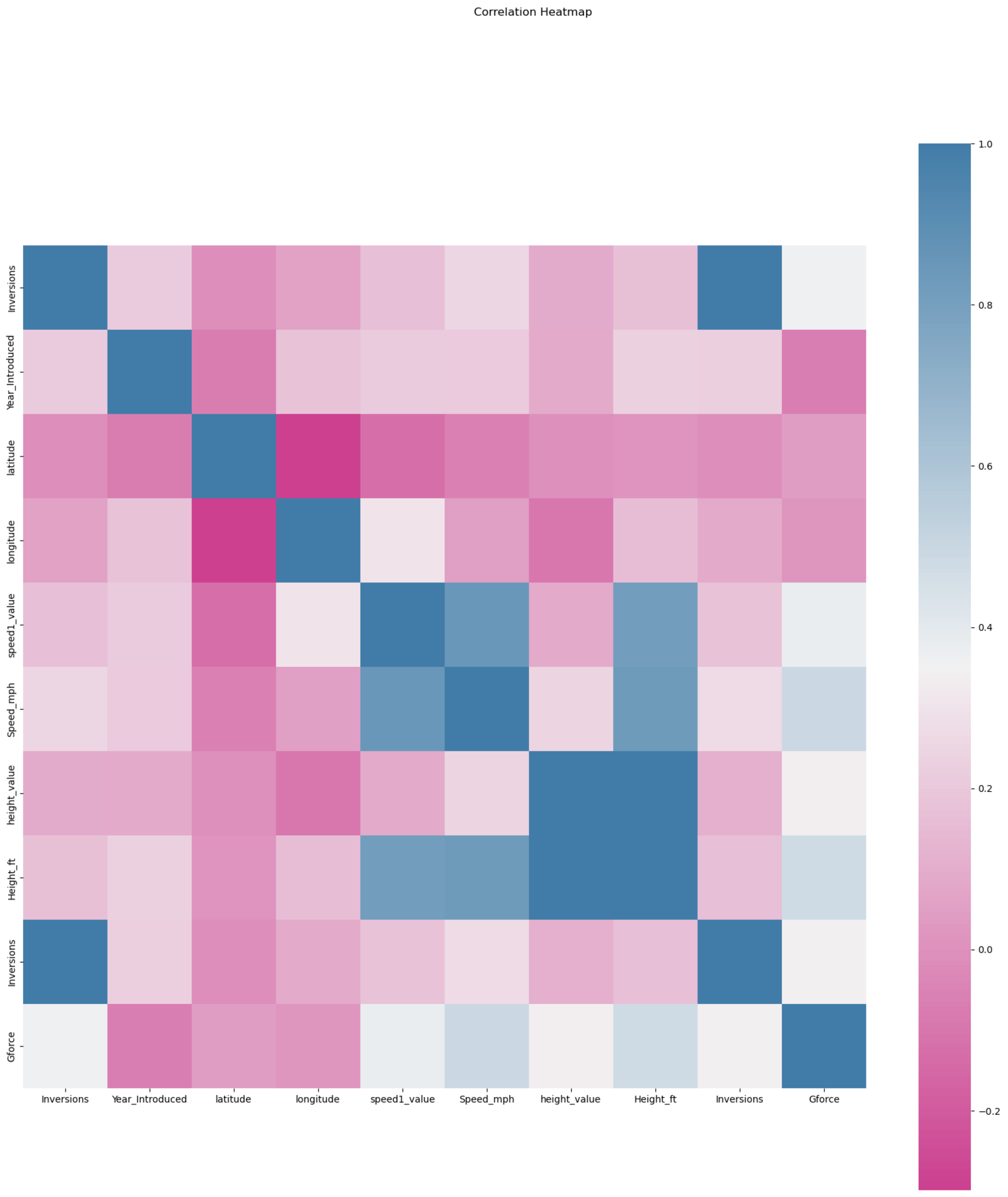
Out[15]:

	Inversions	Year_Introduced	latitude	longitude	speed1_value	Speed_mph	height_value	H
Inversions	1.000000	0.211003	-0.009815	0.061589	0.163419	0.252209	0.094811	0
Year_Introduced	0.211003	1.000000	-0.070982	0.175913	0.210191	0.204853	0.087687	0
latitude	-0.009815	-0.070982	1.000000	-0.298488	-0.121847	-0.063757	-0.004265	0
longitude	0.061589	0.175913	-0.298488	1.000000	0.301179	0.051063	-0.092764	0
speed1_value	0.163419	0.210191	-0.121847	0.301179	1.000000	0.851667	0.088761	0
Speed_mph	0.252209	0.204853	-0.063757	0.051063	0.851667	1.000000	0.241461	0
height_value	0.094811	0.087687	-0.004265	-0.092764	0.088761	0.241461	1.000000	1
Height_ft	0.171330	0.232150	0.011492	0.159733	0.815103	0.829404	1.000000	1
Inversions	1.000000	0.228758	-0.014043	0.087160	0.176105	0.265763	0.108199	0
Gforce	0.356865	-0.066657	0.042871	0.016485	0.379962	0.489337	0.337386	0

In [16]:

```
fig, ax = plt.subplots(figsize=(20,20))
sns.heatmap(correlation,
            cmap=sns.diverging_palette(700,600, as_cmap=True),
            square=True,
            ax=ax)
```

```
fig.suptitle('Correlation Heatmap')
plt.show()
```



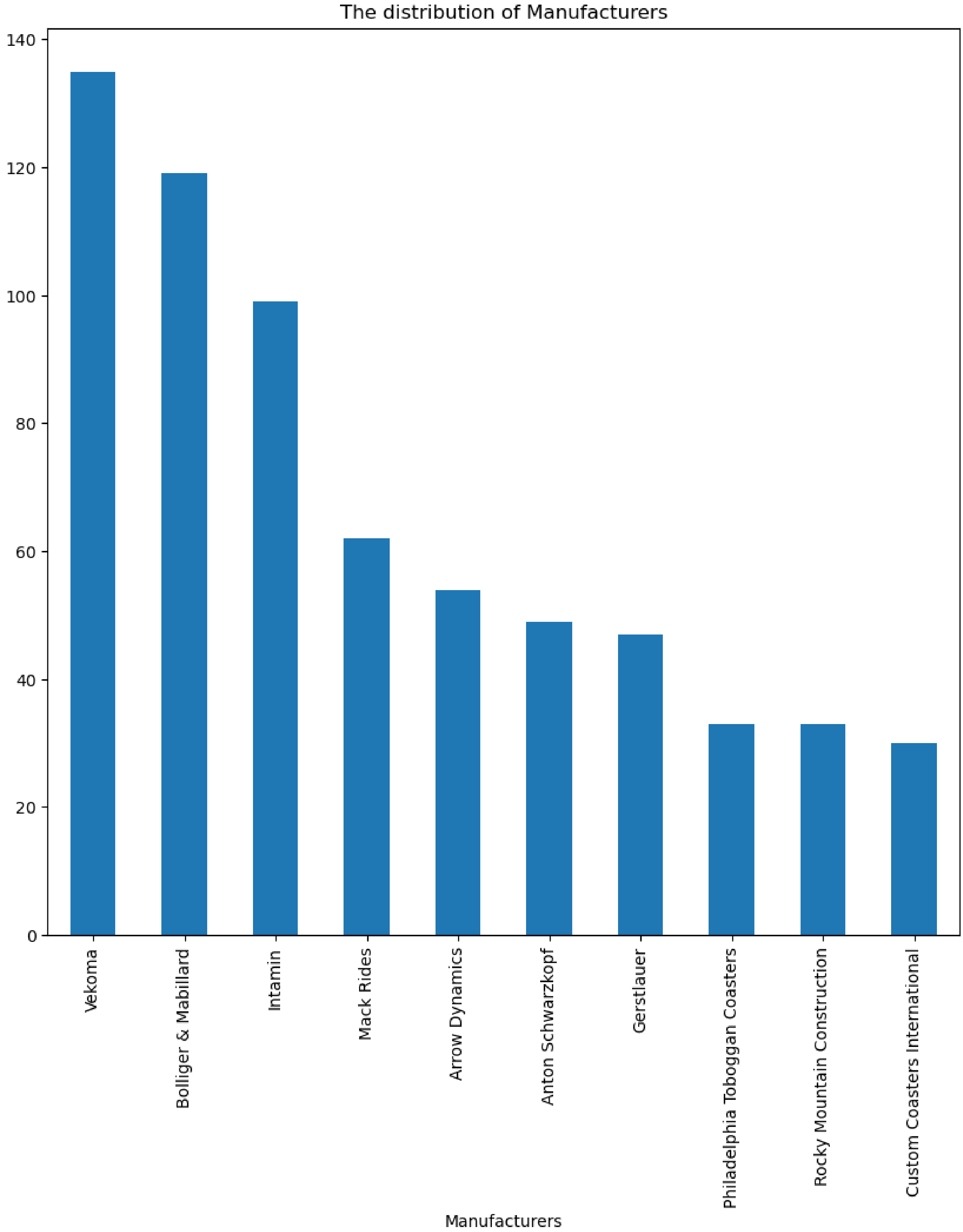
Q9. Optional question: The distribution of the 10 of Manufacturers

In [17]:

```
top10=df.Manufacturer.value_counts()[:10].index.values
x=df[df['Manufacturer'].isin(top10)].Manufacturer.value_counts()
x.plot(kind='bar', figsize=(10,10))
plt.xlabel('Manufacturers')
plt.title('The distribution of Manufacturers')
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

```
Out[17]:
```


The distribution of Manufacturers
140
120
100
80
60
40
20
0
Vekoma
Bolliger & Mabillard
Intamin
Mack Rides
Arrow Dynamics
Anton Schwarzkopf
Gerstlauer
Philadelphia Toboggan Coasters
Rocky Mountain Construction
Custom Coasters International
Manufacturers