In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pylab as plt
import seaborn as sns
plt.style.use('ggplot')
```

In [2]:
```
df = pd.read_csv('coaster_db.csv')
```

In [3]:
```
pd.set_option('display.max_columns', 500)
print(pd.get_option("display.max_columns"))
```

```
500
```

Step 1 : Data Understanding

- Dataframe Shape
- head and tail
- dtypes
- describe

In [4]:
```
df.shape
```

Out[4]:
```
(1087, 56)
```

In [5]:
```
df.head(5)
```

Out[5]:

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

In [6]:
```
df.tail(5)
```

Out[6]:

	coaster_name	Length	Speed	Location	Status	Opening date	Type	M
1082	American Dreier Looping	3,444 ft (1,050 m)	53 mph (85 km/h)	Other	NaN	NaN	Steel	
1083	Pantheon (roller coaster)	3,328 ft (1,014 m)	73 mph (117 km/h)	Busch Gardens Williamsburg	Under construction	2022	Steel – Launched	
1084	Tron Lightcycle Power Run	3,169.3 ft (966.0 m)	59.3[1] mph (95.4 km/h)	Other	NaN	June 16, 2016	Steel – Launched	
1085	Tumbili	770 ft (230 m)	34 mph (55 km/h)	Kings Dominion	Under construction	NaN	Steel – 4th Dimension – Wing Coaster	
1086	Wonder Woman Flight of Courage	3,300 ft (1,000 m)	58 mph (93 km/h)	Six Flags Magic Mountain	Under construction	2022	Steel – Single-rail	

In [7]:

```
df.info()
df.dtypes
```

```
<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
```

```
Out[7]: coaster_name                      object
        Length                            object
        Speed                             object
        Location                          object
        Status                            object
        Opening date                      object
        Type                              object
        Manufacturer                      object
        Height restriction                object
        Model                             object
        Height                            object
        Inversions                       float64
        Lift/launch system                object
        Cost                              object
        Trains                            object
        Park section                      object
        Duration                          object
        Capacity                          object
        G-force                           object
        Designer                          object
        Max vertical angle                object
        Drop                              object
        Soft opening date                 object
        Fast Lane available               object
        Replaced                          object
        Track layout                      object
        Fastrack available                object
        Soft opening date.1               object
        Closing date                      object
        Opened                            object
        Replaced by                       object
        Website                           object
        Flash Pass Available              object
        Must transfer from wheelchair     object
        Theme                             object
        Single rider line available       object
        Restraint Style                   object
        Flash Pass available              object
        Acceleration                      object
        Restraints                        object
        Name                              object
        year_introduced                    int64
        latitude                         float64
        longitude                        float64
        Type_Main                         object
        opening_date_clean                object
        speed1                            object
        speed2                            object
        speed1_value                     float64
        speed1_unit                       object
        speed_mph                        float64
        height_value                     float64
        height_unit                       object
        height_ft                        float64
        Inversions_clean                   int64
        Gforce_clean                     float64
        dtype: object
```

In [8]:
```
df.describe()
```

Out[8]:

	Inversions	year_introduced	latitude	longitude	speed1_value	speed_mph	height
count	932.000000	1087.000000	812.000000	812.000000	937.000000	937.000000	965.0
mean	1.547210	1994.986201	38.373484	-41.595373	53.850374	48.617289	89.5
std	2.114073	23.475248	15.516596	72.285227	23.385518	16.678031	136.2
min	0.000000	1884.000000	-48.261700	-123.035700	5.000000	5.000000	4.0
25%	0.000000	1989.000000	35.031050	-84.552200	40.000000	37.300000	44.0
50%	0.000000	2000.000000	40.289800	-76.653600	50.000000	49.700000	79.0
75%	3.000000	2010.000000	44.799600	2.778100	63.000000	58.000000	113.0
max	14.000000	2022.000000	63.230900	153.426500	240.000000	149.100000	3937.0

Step 2 : Data Preparation

- Dropping irrelavent columns ans rows
- Identifying duplicated columns
- Renaming columns
- Feature creation

In [9]:
```
df.columns
```

Out[9]:
```
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_clean', 'speed1', 'speed2', 'speed1_value', 'speed1_unit',
       'speed_mph', 'height_value', 'height_unit', 'height_ft',
       'Inversions_clean', 'Gforce_clean'],
      dtype='object')
```

In [10]:
```
# remove unnecessary columns
df = df[['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_cl
    #'speed1', 'speed2', 'speed1_value', 'speed1_unit',
       'speed_mph',
    #'height_value', 'height_unit',
    'height_ft',  'Inversions_clean', 'Gforce_clean']].copy()
```

In [11]:

```
df.dtypes
```

Out[11]:

```
coaster_name           object
Location               object
Status                 object
Manufacturer           object
year_introduced         int64
latitude              float64
longitude             float64
Type_Main              object
opening_date_clean     object
speed_mph             float64
height_ft             float64
Inversions_clean        int64
Gforce_clean          float64
dtype: object
```

In [12]:

```
#change datatypes
df['opening_date_clean'] = pd.to_datetime(df['opening_date_clean'])
df['opening_date_clean']
```

Out[12]:

```
0      1884-06-16
1      1895-01-01
2             NaT
3      1901-01-01
4      1901-01-01
          ...
1082          NaT
1083   2022-01-01
1084   2016-06-16
1085          NaT
1086   2022-01-01
Name: opening_date_clean, Length: 1087, dtype: datetime64[ns]
```

In [13]:

```
#Rename Columns
df.columns
```

Out[13]:

```
Index(['coaster_name', 'Location', 'Status', 'Manufacturer', 'year_introduced',
       'latitude', 'longitude', 'Type_Main', 'opening_date_clean', 'speed_mph',
       'height_ft', 'Inversions_clean', 'Gforce_clean'],
      dtype='object')
```

In [14]:

```
df = df.rename(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 [15]:

```
df.head(5)
```

Out[15]:

	Coaster_Name	Location	Status	Manufacturer	Year_Introduced	latitude	longitude	Ty
0	Switchback Railway	Coney Island	Removed	LaMarcus Adna Thompson	1884	40.5740	-73.9780	
1	Flip Flap Railway	Sea Lion Park	Removed	Lina Beecher	1895	40.5780	-73.9790	
2	Switchback Railway (Euclid Beach Park)	Cleveland, Ohio, United States	Closed	NaN	1896	41.5800	-81.5700	
3	Loop the Loop (Coney Island)	Other	Removed	Edwin Prescott	1901	40.5745	-73.9780	
4	Loop the Loop (Young's Pier)	Other	Removed	Edwin Prescott	1901	39.3538	-74.4342	

In [16]:

```
#cleaning null values
df.isna().sum()
```

Out[16]:

```
Coaster_Name         0
Location             0
Status             213
Manufacturer        59
Year_Introduced      0
latitude           275
longitude          275
Type_Main            0
Opening_Date       250
Speed_mph          150
Height_ft          916
Inversions           0
Gforce             725
dtype: int64
```

In [17]:

```
#remove duplicated values
df.loc[df.duplicated()]
```

Out[17]:

Coaster_Name	Location	Status	Manufacturer	Year_Introduced	latitude	longitude	Type_M

In [18]:

```
df.loc[df.duplicated(subset=['Coaster_Name'])]
```

Out[18]:

	Coaster_Name	Location	Status	Manufacturer	Year_Introduced	latitude	longit
43	Crystal Beach Cyclone	Crystal Beach Park	Removed	Traver Engineering	1927	42.8617	-79.0
60	Derby Racer	Revere Beach	Removed	Fred W. Pearce	1937	42.4200	-70.9
61	Blue Streak (Conneaut Lake)	Conneaut Lake Park	Closed	NaN	1938	41.6349	-80.3
167	Big Thunder Mountain Railroad	Other	NaN	Arrow Development (California and Florida)Dyna...	1980	NaN	N
237	Thunder Run (Canada's Wonderland)	Canada's Wonderland	Operating	Mack Rides	1986	43.8427	-79.5
...	...	...	...	...	...	...	
1063	Lil' Devil Coaster	Six Flags Great Adventure	Operating	Zamperla	2021	40.1343	-74.4
1064	Little Dipper (Conneaut Lake Park)	Conneaut Lake Park	Operating	Allan Herschell Company	2021	41.6343	-80.3
1080	Iron Gwazi	Busch Gardens Tampa Bay	Under construction	Rocky Mountain Construction	2022	28.0339	-82.4
1082	American Dreier Looping	Other	NaN	Anton Schwarzkopf	2022	NaN	N
1084	Tron Lightcycle Power Run	Other	NaN	Vekoma	2022	NaN	N

97 rows × 13 columns

In [19]:
```
df.query('Coaster_Name =="Tron Lightcycle Power Run"')
```

Out[19]:

	Coaster_Name	Location	Status	Manufacturer	Year_Introduced	latitude	longitude	Ty
978	Tron Lightcycle Power Run	Other	NaN	Vekoma	2016	NaN	NaN	
1084	Tron Lightcycle Power Run	Other	NaN	Vekoma	2022	NaN	NaN	

In [20]:
```
df.duplicated(subset=['Coaster_Name','Location','Opening_Date']).sum()
```

Out[20]:
```
97
```

In [21]:
```
df = df.loc[~df.duplicated(subset=['Coaster_Name','Location','Opening_Date'])]\
.reset_index(drop=True).copy()
```

In [22]:
```
df
```

Out[22]:

	Coaster_Name	Location	Status	Manufacturer	Year_Introduced	latitude	longit
0	Switchback Railway	Coney Island	Removed	LaMarcus Adna Thompson	1884	40.5740	-73.9
1	Flip Flap Railway	Sea Lion Park	Removed	Lina Beecher	1895	40.5780	-73.9
2	Switchback Railway (Euclid Beach Park)	Cleveland, Ohio, United States	Closed	NaN	1896	41.5800	-81.5
3	Loop the Loop (Coney Island)	Other	Removed	Edwin Prescott	1901	40.5745	-73.9
4	Loop the Loop (Young's Pier)	Other	Removed	Edwin Prescott	1901	39.3538	-74.4
...	...	...	...	...	...	...	
985	Ice Breaker (roller coaster)	SeaWorld Orlando	Under construction	Premier Rides	2022	28.4088	-81.4
986	Leviathan (Sea World)	Sea World	Under construction	Martin & Vleminckx	2022	-27.9574	153.4
987	Pantheon (roller coaster)	Busch Gardens Williamsburg	Under construction	Intamin	2022	37.2339	-76.6
988	Tumbili	Kings Dominion	Under construction	S&S – Sansei Technologies	2022	NaN	N
989	Wonder Woman Flight of Courage	Six Flags Magic Mountain	Under construction	Rocky Mountain Construction	2022	NaN	N

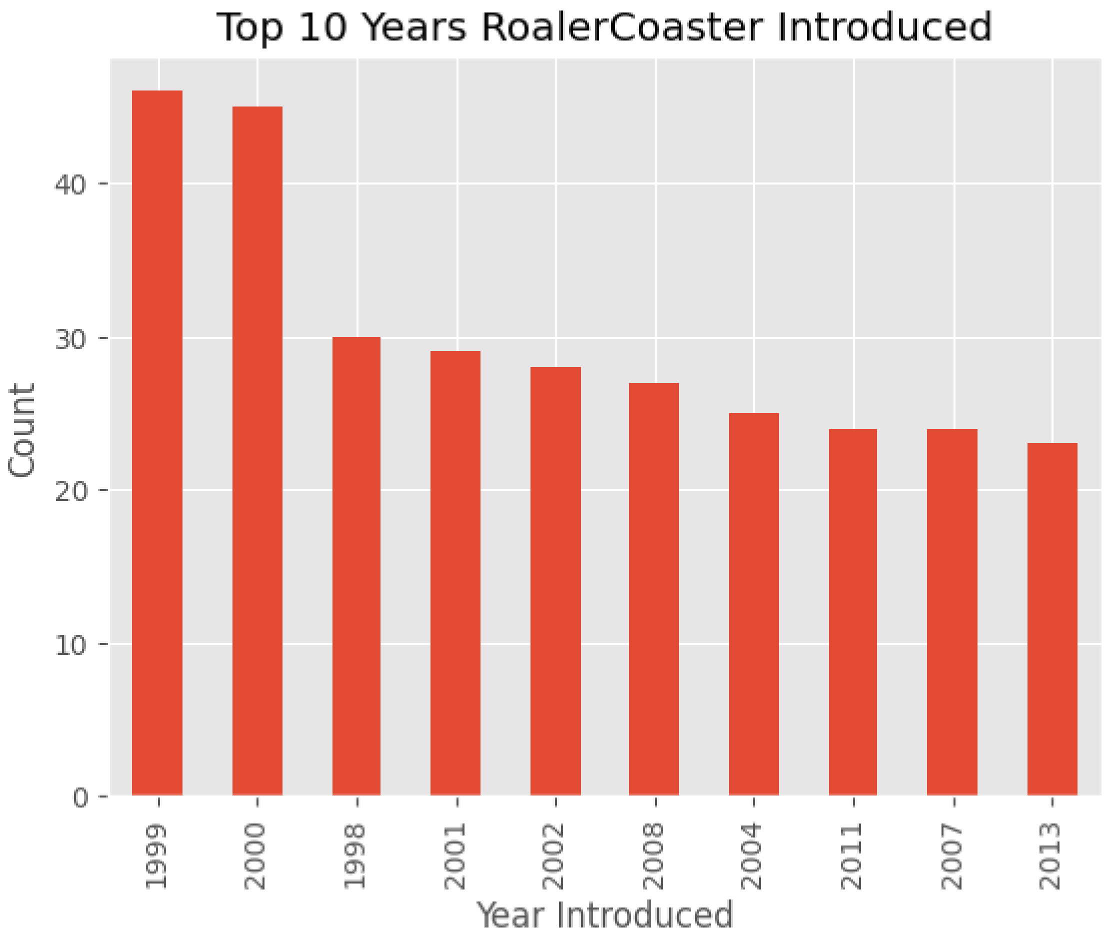
990 rows × 13 columns

Step 3 : Feature Understandings

Univariate Analysis

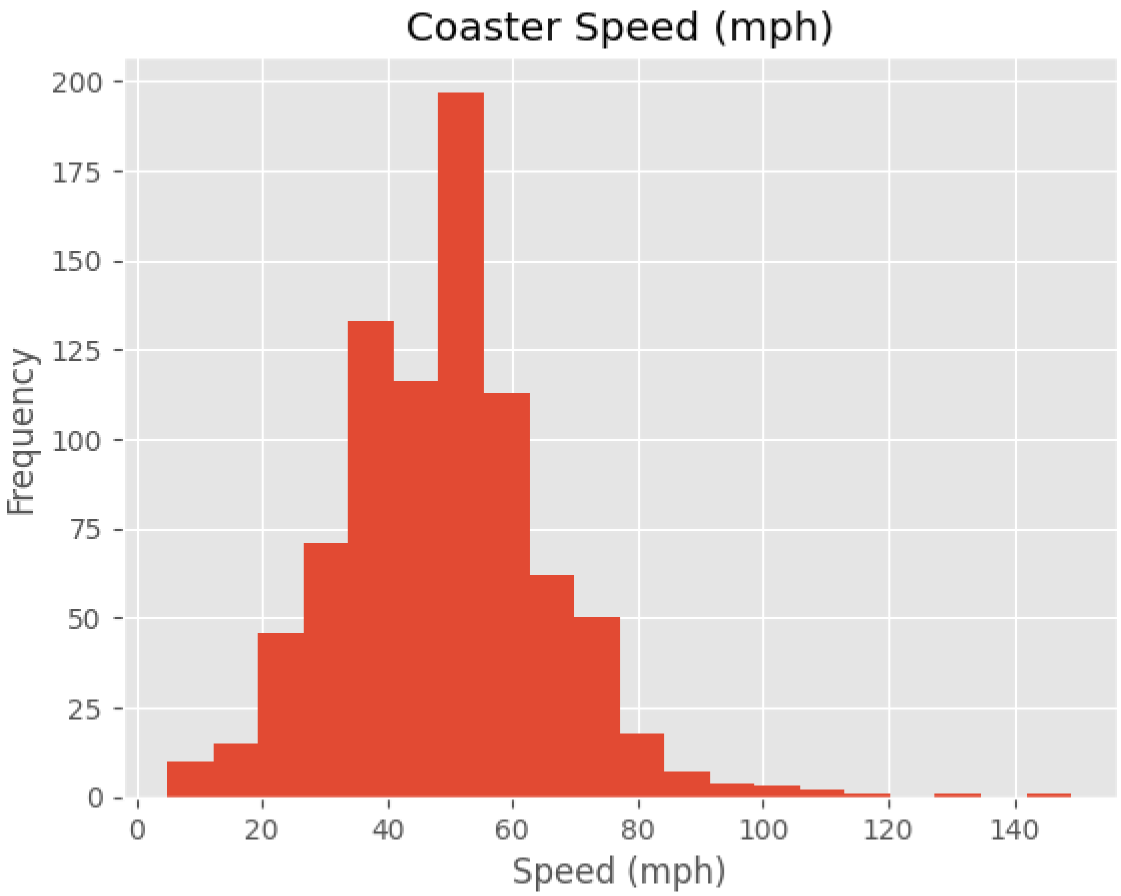
- Plotting Feature Distributions
  - Histograms
  - KDE
  - Boxplot

In [23]:
```
ax = df['Year_Introduced'].value_counts().head(10) \
     .plot(kind='bar', title='Top 10 Years RoalerCoaster Introduced')
ax.set_xlabel('Year Introduced')
ax.set_ylabel('Count')
plt.show()
```



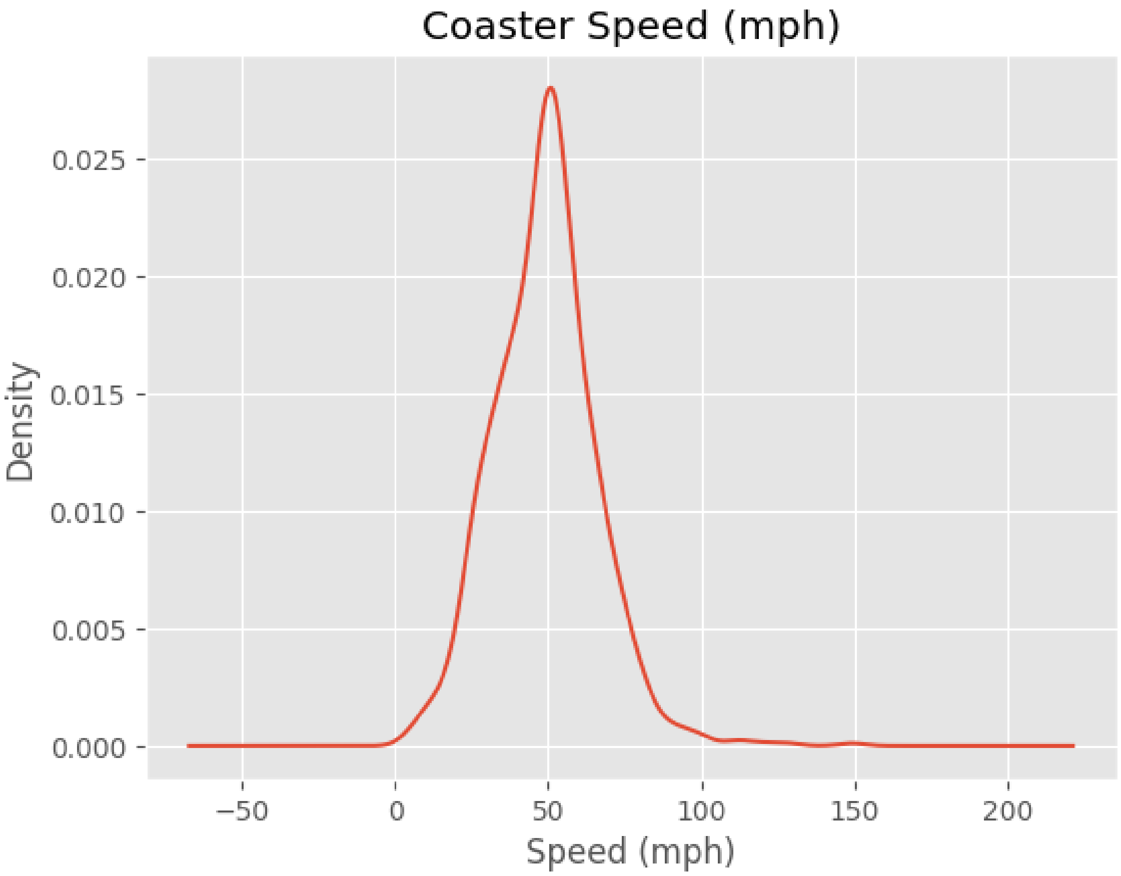
In [24]:

```
ax = df['Speed_mph'].plot(kind='hist',
                          bins=20,
                          title='Coaster Speed (mph)')
ax.set_xlabel('Speed (mph)')
plt.show()
```



In [25]:

```
ax = df['Speed_mph'].plot(kind='kde',
                          title='Coaster Speed (mph)')
ax.set_xlabel('Speed (mph)')
plt.show()
```

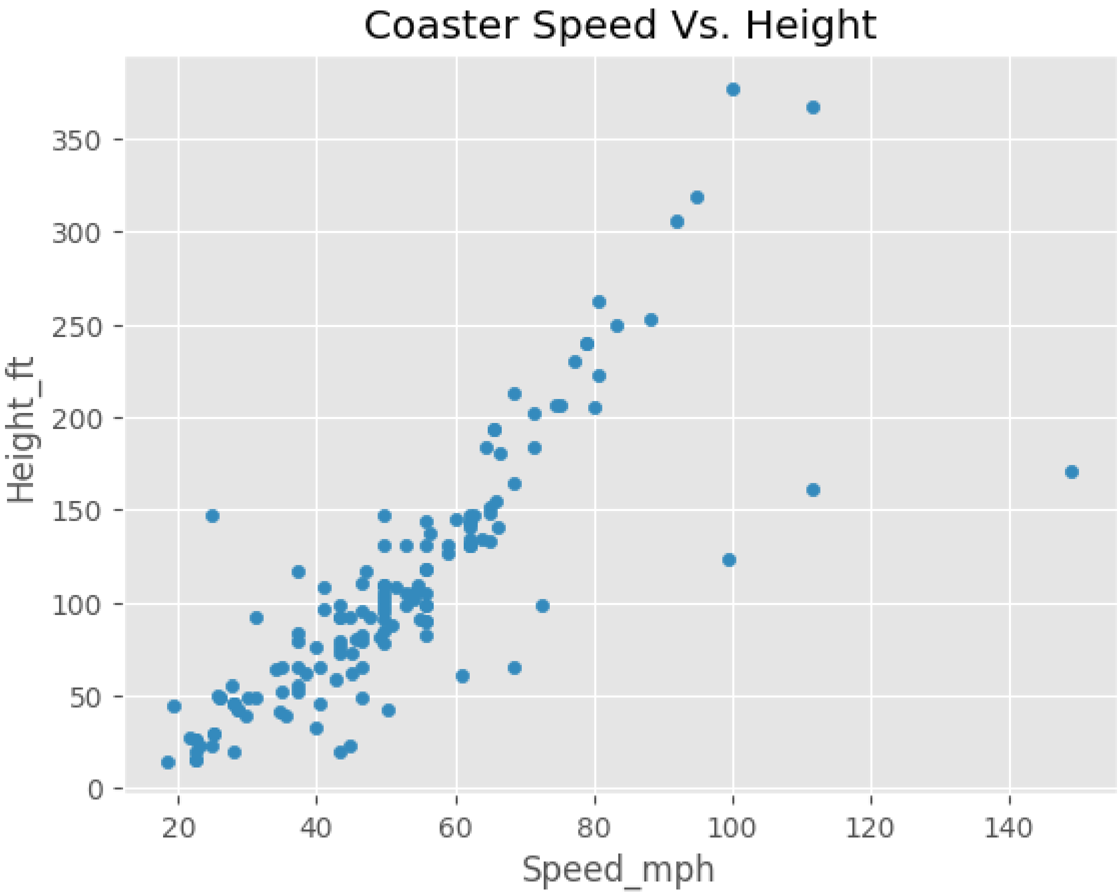


Step 4 : Feature Relationships

- ScatterPlot
- Heatmap Correlation
- Pairplot
- Groupby Comparisons

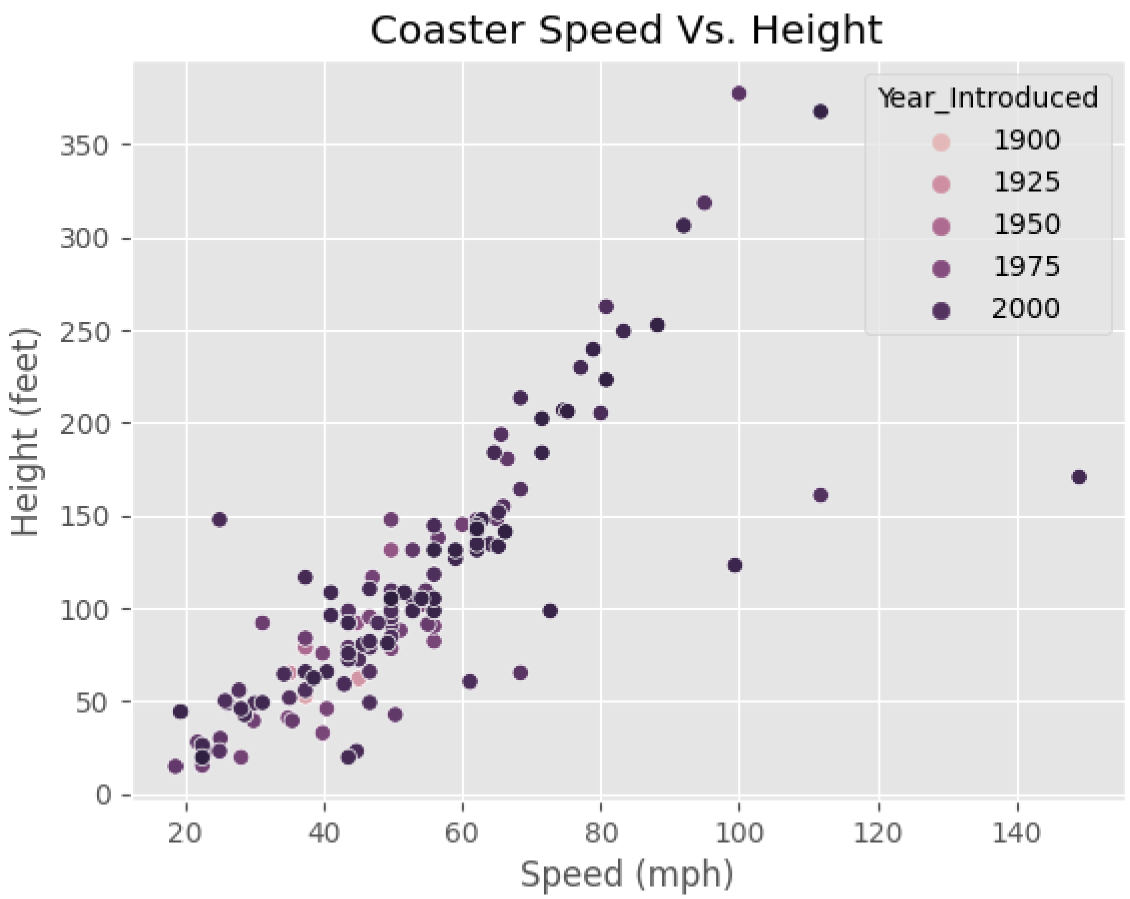
In [26]:

```
df.plot(kind='scatter',
        x='Speed_mph',
        y='Height_ft',
        title='Coaster Speed Vs. Height')
plt.show()
```

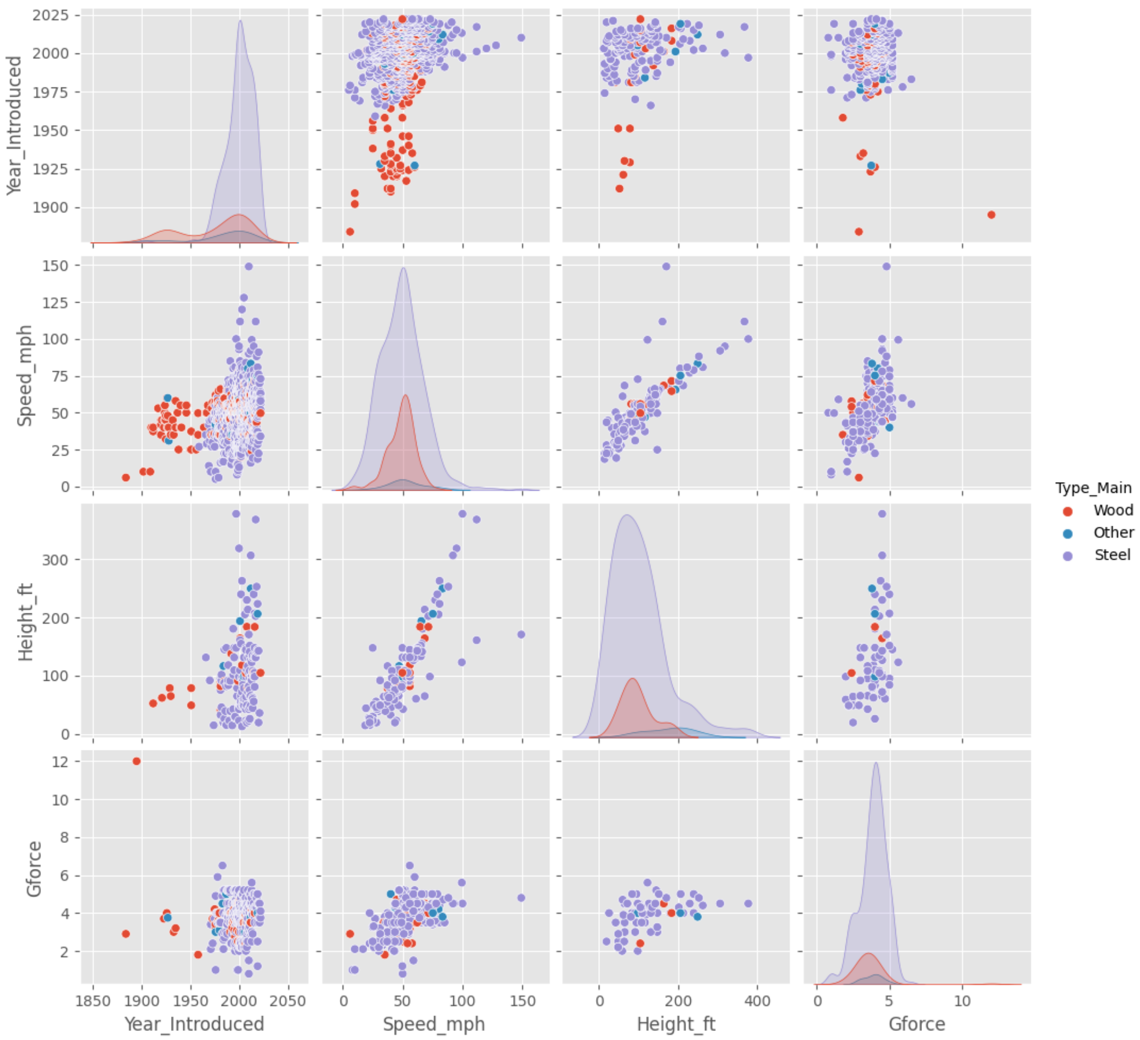


In [27]:

```
ax = sns.scatterplot(data=df,
                x='Speed_mph',
                y='Height_ft',
                hue='Year_Introduced')
ax.set_title('Coaster Speed Vs. Height')
ax.set_xlabel('Speed (mph)')
ax.set_ylabel('Height (feet)')
plt.show()
```



```
In [28]: sns.pairplot(df,
                      vars=['Year_Introduced','Speed_mph','Height_ft','Gforce'],
                      hue='Type_Main')
         plt.show()
```



In [29]:

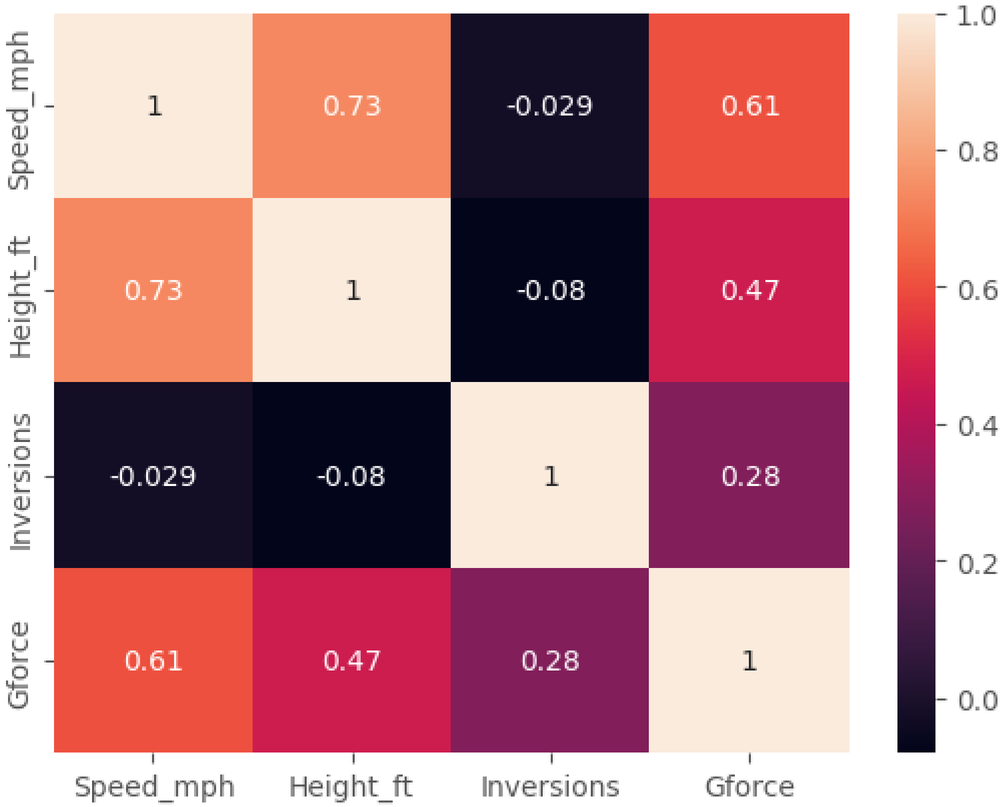
```
df_corr = df[['Speed_mph','Height_ft','Inversions','Gforce']].dropna().corr()
df_corr
```

Out[29]:

	Speed_mph	Height_ft	Inversions	Gforce
Speed_mph	1.000000	0.733999	-0.028705	0.607383
Height_ft	0.733999	1.000000	-0.079736	0.466482
Inversions	-0.028705	-0.079736	1.000000	0.275991
Gforce	0.607383	0.466482	0.275991	1.000000

In [30]:

```
sns.heatmap(df_corr, annot=True)
plt.show()
```



Step 4 : Ask a question about the data

- Try to answer a question you have about data using a plot or statistic.

1. What are the top 10 locations, with the faster roller coaster ?

In [31]:

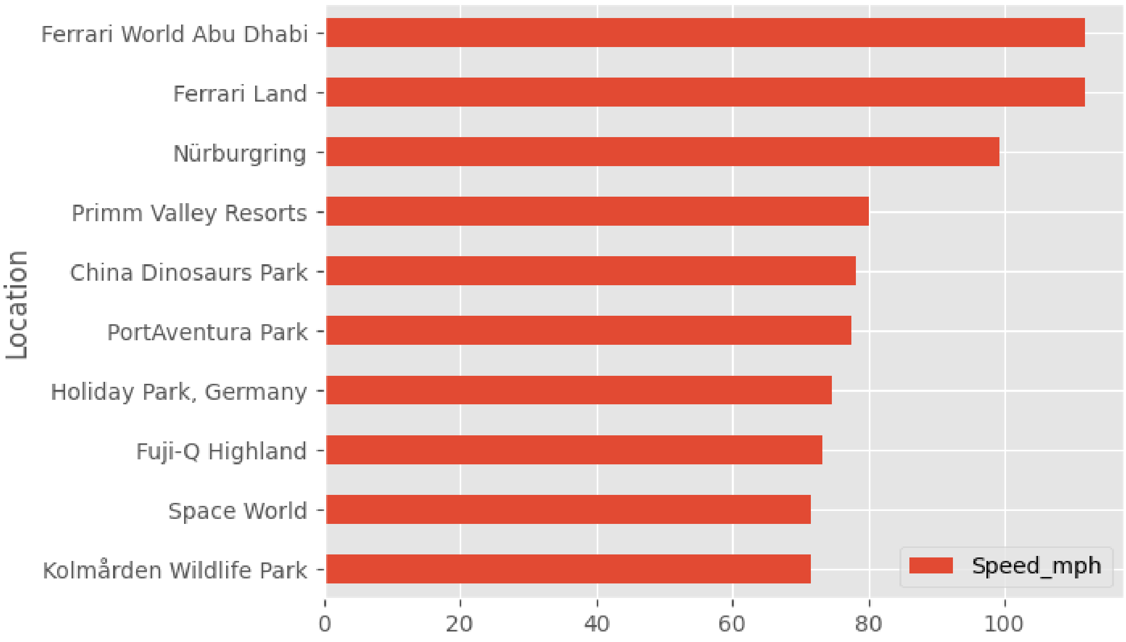
```
df_loc_speed = df[['Location','Speed_mph']].dropna().groupby('Location').mean()
df1 = df_loc_speed.sort_values(by=['Speed_mph'], ascending=False).head(10)
df1
```

Out[31]:

Location	Speed_mph
Ferrari World Abu Dhabi	111.850000
Ferrari Land	111.800000
Nürburgring	99.400000
Primm Valley Resorts	80.000000
China Dinosaurs Park	78.300000
PortAventura Park	77.400000
Holiday Park, Germany	74.600000
Fuji-Q Highland	73.133333
Kolmården Wildlife Park	71.500000
Space World	71.500000

In [32]:

```
df1.sort_values(by=['Speed_mph']).plot(kind='barh')
plt.show()
```



2. What are the top 10 locations, with the faster roller coaster (minimum 10 coaster at single location) ?

In [33]:

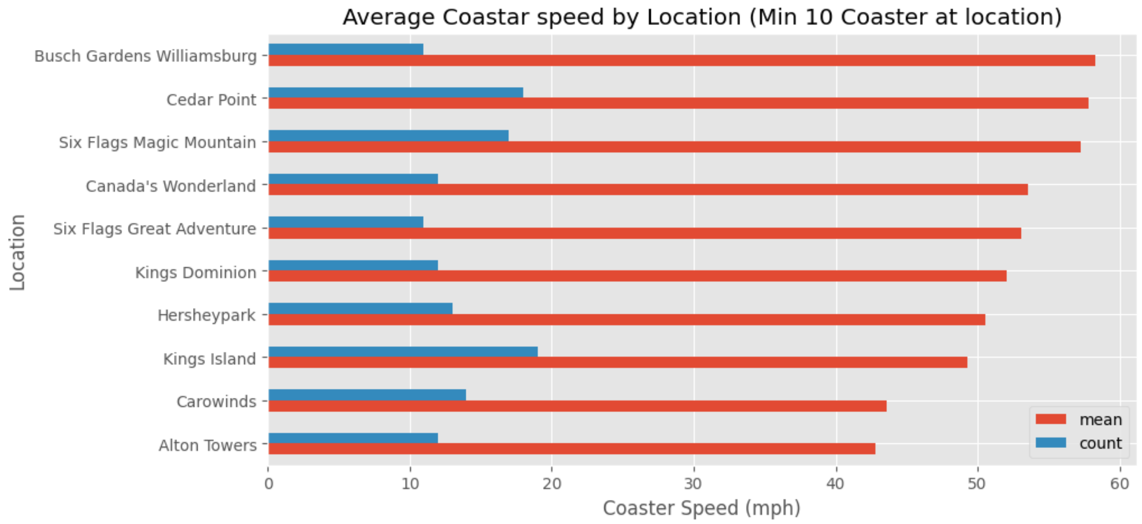
```
df2 = df.query('Location != "Other"') \
  .groupby('Location')['Speed_mph'] \
  .aggregate(['mean','count']) \
  .query('count >= 10') \
  .sort_values('mean')
df2
```

Out[33]:

	mean	count
Location		
Alton Towers	42.791667	12
Carowinds	43.571429	14
Kings Island	49.273684	19
Hersheypark	50.576923	13
Kings Dominion	52.083333	12
Six Flags Great Adventure	53.036364	11
Canada's Wonderland	53.533333	12
Six Flags Magic Mountain	57.241176	17
Cedar Point	57.833333	18
Busch Gardens Williamsburg	58.318182	11

In [34]:

```python
ax = df2.plot(kind='barh',
        title='Average Coastar speed by Location (Min 10 Coaster at location)',
        figsize=(10,5)
       )
ax.set_xlabel('Coaster Speed (mph)')
plt.show()
```



In [ ]: