In [257]:

```
import pandas as pd
import numpy as np
import matplotlib.pylab as plt
import seaborn as sns
plt.style.use('ggplot')
pd.set_option('display.max_columns',200)
```

In [258]:

```
df = pd.read_csv(r'D:\Data+Science@Consoleflare\Pandas\RollerCoaster\coaster_db.csv')
```

Data Understanding

In [259]:

```
df.shape
```

Out[259]:

```
(1087, 56)
```

In [260]:

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

Out[260]:

	coaster_name	Length	Speed	Location	Status	Opening date	Type	Manufacturer	Height restriction	Model	H
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	

Data preparation

In [261]:

```
df.columns
```

Out[261]:

```
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 [262]:

```
df.dtypes
```

Out[262]:

```
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
```

In [263]:

```
df.describe()
```

Out[263]:

	Inversions	year_introduced	latitude	longitude	speed1_value	speed_mph	height_value	height_f
count	932.000000	1087.000000	812.000000	812.000000	937.000000	937.000000	965.000000	171.000000
mean	1.547210	1994.986201	38.373484	-41.595373	53.850374	48.617289	89.575171	101.996491
std	2.114073	23.475248	15.516596	72.285227	23.385518	16.678031	136.246444	67.329092
min	0.000000	1884.000000	-48.261700	-123.035700	5.000000	5.000000	4.000000	13.100000
25%	0.000000	1989.000000	35.031050	-84.552200	40.000000	37.300000	44.000000	51.800000
50%	0.000000	2000.000000	40.289800	-76.653600	50.000000	49.700000	79.000000	91.200000
75%	3.000000	2010.000000	44.799600	2.778100	63.000000	58.000000	113.000000	131.200000
max	14.000000	2022.000000	63.230900	153.426500	240.000000	149.100000	3937.000000	377.300000

In [264]:

```
#Example of dropping column
#df.drop(['Opening date''],axis=1)
```

In [265]:

```
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_clean',
    #'speed1', 'speed2', 'speed1_value', 'speed1_unit',
       'speed_mph',
    #'height_value', 'height_unit',
    'height_ft',
       'Inversions_clean', 'Gforce_clean']].copy()
```

In [266]:

```
df
```

Out[266]:

	coaster_name	Location	Status	Manufacturer	year_introduced	latitude	longitude	Type_Main	ope
0	Switchback Railway	Coney Island	Removed	LaMarcus Adna Thompson	1884	40.5740	-73.9780	Wood	
1	Flip Flap Railway	Sea Lion Park	Removed	Lina Beecher	1895	40.5780	-73.9790	Wood	
2	Switchback Railway (Euclid Beach Park)	Cleveland, Ohio, United States	Closed	NaN	1896	41.5800	-81.5700	Other	
3	Loop the Loop (Coney Island)	Other	Removed	Edwin Prescott	1901	40.5745	-73.9780	Steel	
4	Loop the Loop (Young's Pier)	Other	Removed	Edwin Prescott	1901	39.3538	-74.4342	Steel	
...	...	...	...	...	...	...	...	...	
1082	American Dreier Looping	Other	NaN	Anton Schwarzkopf	2022	NaN	NaN	Steel	
1083	Pantheon (roller coaster)	Busch Gardens Williamsburg	Under construction	Intamin	2022	37.2339	-76.6426	Steel	
1084	Tron Lightcycle Power Run	Other	NaN	Vekoma	2022	NaN	NaN	Steel	
1085	Tumbili	Kings Dominion	Under construction	S&S – Sansei Technologies	2022	NaN	NaN	Steel	
1086	Wonder Woman Flight of Courage	Six Flags Magic Mountain	Under construction	Rocky Mountain Construction	2022	NaN	NaN	Steel	

1087 rows × 13 columns

In [267]:

```
df.dtypes
```

Out[267]:

```
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 [268]:

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

In [269]:

```
#Rename columns
df = df.rename(columns={'coaster_name':'Coaster_Name','opening_date_clean':'Opening_Date','year_introduce
                        'speed_mph':'Speed_mph','height_ft':'Height_ft','Inversions_clean':'Inversions','Gforce
df
```

Out[269]:

	Coaster_Name	Location	Status	Manufacturer	Year_Introduced	latitude	longitude	Type_Main	Op
0	Switchback Railway	Coney Island	Removed	LaMarcus Adna Thompson	1884	40.5740	-73.9780	Wood	
1	Flip Flap Railway	Sea Lion Park	Removed	Lina Beecher	1895	40.5780	-73.9790	Wood	
2	Switchback Railway (Euclid Beach Park)	Cleveland, Ohio, United States	Closed	NaN	1896	41.5800	-81.5700	Other	
3	Loop the Loop (Coney Island)	Other	Removed	Edwin Prescott	1901	40.5745	-73.9780	Steel	
4	Loop the Loop (Young's Pier)	Other	Removed	Edwin Prescott	1901	39.3538	-74.4342	Steel	
...	...	...	...	...	...	...	...	...	
1082	American Dreier Looping	Other	NaN	Anton Schwarzkopf	2022	NaN	NaN	Steel	
1083	Pantheon (roller coaster)	Busch Gardens Williamsburg	Under construction	Intamin	2022	37.2339	-76.6426	Steel	
1084	Tron Lightcycle Power Run	Other	NaN	Vekoma	2022	NaN	NaN	Steel	
1085	Tumbili	Kings Dominion	Under construction	S&S – Sansei Technologies	2022	NaN	NaN	Steel	
1086	Wonder Woman Flight of Courage	Six Flags Magic Mountain	Under construction	Rocky Mountain Construction	2022	NaN	NaN	Steel	

1087 rows × 13 columns

In [270]:

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

Out[270]:

```
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 [271]:

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

Out[271]:

	Coaster_Name	Location	Status	Manufacturer	Year_Introduced	latitude	longitude	Type_Main	Op
43	Crystal Beach Cyclone	Crystal Beach Park	Removed	Traver Engineering	1927	42.8617	-79.0598	Wood	
60	Derby Racer	Revere Beach	Removed	Fred W. Pearce	1937	42.4200	-70.9860	Wood	
61	Blue Streak (Conneaut Lake)	Conneaut Lake Park	Closed	NaN	1938	41.6349	-80.3180	Wood	
167	Big Thunder Mountain Railroad	Other	NaN	Arrow Development (California and Florida)Dyna...	1980	NaN	NaN	Steel	
237	Thunder Run (Canada's Wonderland)	Canada's Wonderland	Operating	Mack Rides	1986	43.8427	-79.5423	Steel	
...	...	...	...	...	...	...	...	...	
1063	Lil' Devil Coaster	Six Flags Great Adventure	Operating	Zamperla	2021	40.1343	-74.4434	Steel	
1064	Little Dipper (Conneaut Lake Park)	Conneaut Lake Park	Operating	Allan Herschell Company	2021	41.6343	-80.3165	Steel	
1080	Iron Gwazi	Busch Gardens Tampa Bay	Under construction	Rocky Mountain Construction	2022	28.0339	-82.4231	Steel	
1082	American Dreier Looping	Other	NaN	Anton Schwarzkopf	2022	NaN	NaN	Steel	
1084	Tron Lightcycle Power Run	Other	NaN	Vekoma	2022	NaN	NaN	Steel	

97 rows × 13 columns

In [272]:

```
#checking an example of duplicate
df.query("Coaster_Name == 'Iron Gwazi'")
```

Out[272]:

	Coaster_Name	Location	Status	Manufacturer	Year_Introduced	latitude	longitude	Type_Main	Openi
482	Iron Gwazi	Busch Gardens Tampa Bay	Under construction	Rocky Mountain Construction	1999	28.0339	-82.4231	Steel	
1080	Iron Gwazi	Busch Gardens Tampa Bay	Under construction	Rocky Mountain Construction	2022	28.0339	-82.4231	Steel	

In [273]:

```
df.columns
```

Out[273]:

```
Index(['Coaster_Name', 'Location', 'Status', 'Manufacturer', 'Year_Introduced',
       'latitude', 'longitude', 'Type_Main', 'Opening_Date', 'Speed_mph',
       'Height_ft', 'Inversions', 'Gforce'],
      dtype='object')
```

In [274]:

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

Out[274]:

	Coaster_Name	Location	Status	Manufacturer	Year_Introduced	latitude	longitude	Type_Main	Op
0	Switchback Railway	Coney Island	Removed	LaMarcus Adna Thompson	1884	40.5740	-73.9780	Wood	
1	Flip Flap Railway	Sea Lion Park	Removed	Lina Beecher	1895	40.5780	-73.9790	Wood	
2	Switchback Railway (Euclid Beach Park)	Cleveland, Ohio, United States	Closed	NaN	1896	41.5800	-81.5700	Other	
3	Loop the Loop (Coney Island)	Other	Removed	Edwin Prescott	1901	40.5745	-73.9780	Steel	
4	Loop the Loop (Young's Pier)	Other	Removed	Edwin Prescott	1901	39.3538	-74.4342	Steel	
...	...	...	...	...	...	...	...	...	
985	Ice Breaker (roller coaster)	SeaWorld Orlando	Under construction	Premier Rides	2022	28.4088	-81.4633	Steel	
986	Leviathan (Sea World)	Sea World	Under construction	Martin & Vleminckx	2022	-27.9574	153.4263	Wood	
987	Pantheon (roller coaster)	Busch Gardens Williamsburg	Under construction	Intamin	2022	37.2339	-76.6426	Steel	
988	Tumbili	Kings Dominion	Under construction	S&S – Sansei Technologies	2022	NaN	NaN	Steel	
989	Wonder Woman Flight of Courage	Six Flags Magic Mountain	Under construction	Rocky Mountain Construction	2022	NaN	NaN	Steel	

990 rows × 13 columns

Plotting Feature Understanding

In [275]:

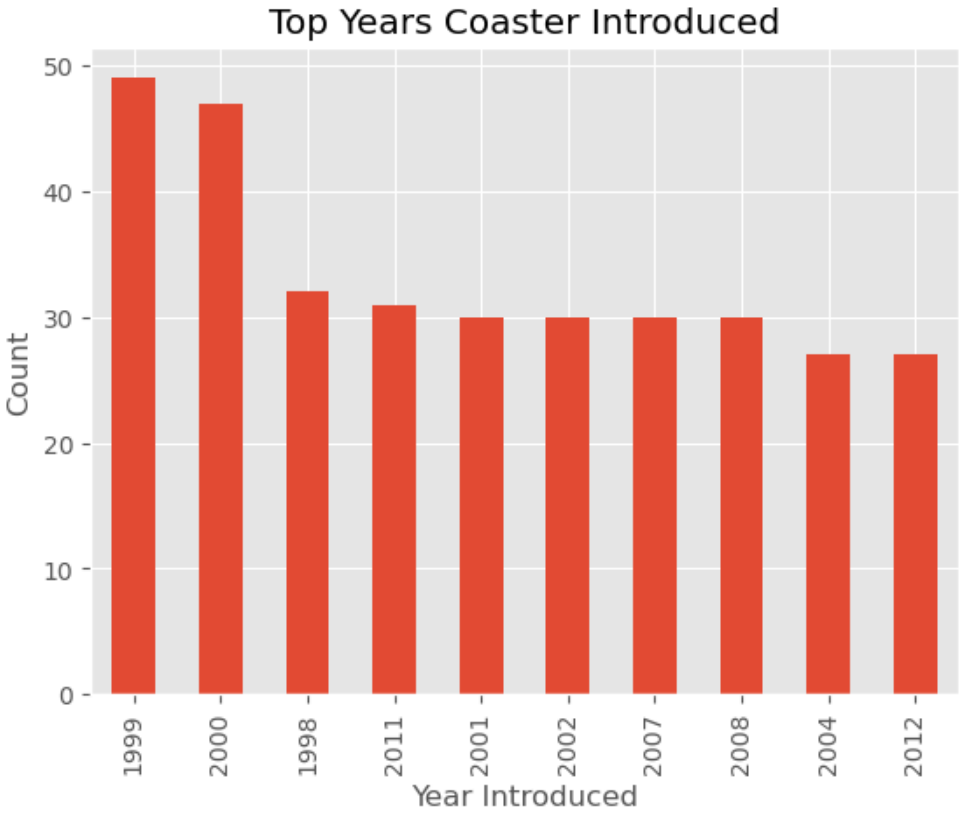
```
#{Univariate Analysis}

##Plotting Feature Distribution
    #Histogram
    #KDE
    #Boxplot
```

In [276]:

```
ax = df['Year_Introduced'].value_counts().head(10).plot(kind = 'bar', title = 'Top Years Coaster Introduc
ax.set_xlabel('Year Introduced')
ax.set_ylabel('Count')
```

Out[276]:

```
Text(0, 0.5, 'Count')
```

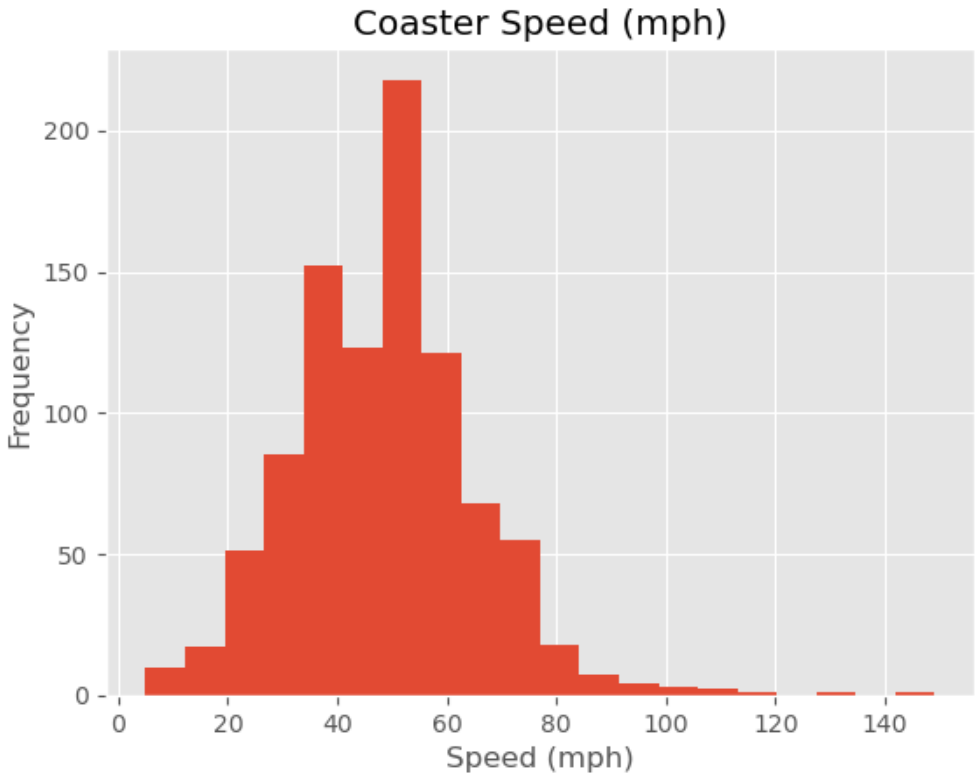


In [277]:

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

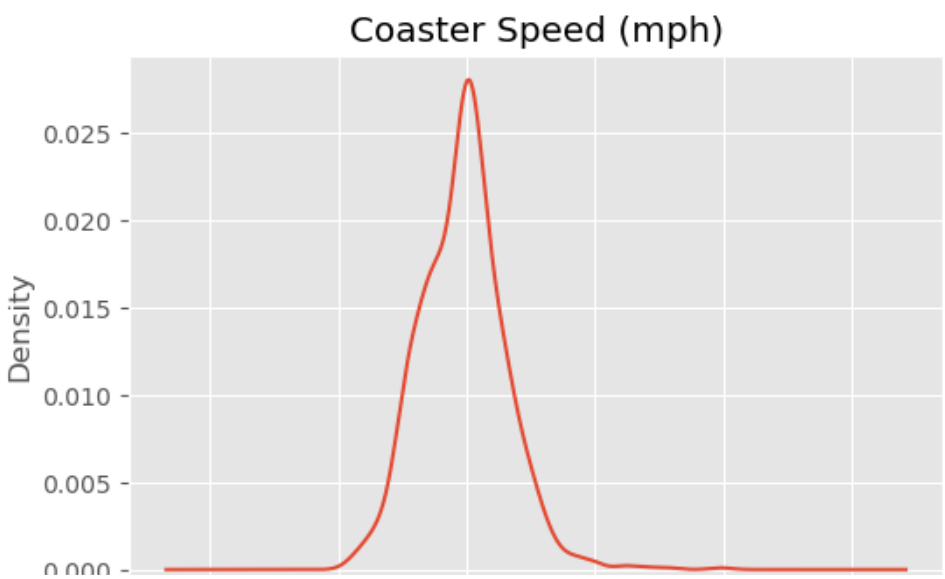
Out[277]:

```
Text(0.5, 0, 'Speed (mph)')
```



In [278]:

```
ax = df['Speed_mph'].plot(kind = 'kde',title = 'Coaster Speed (mph)',figsize = (6,4))
ax.set_xlabel('Speed (mph)')
plt.show()
```



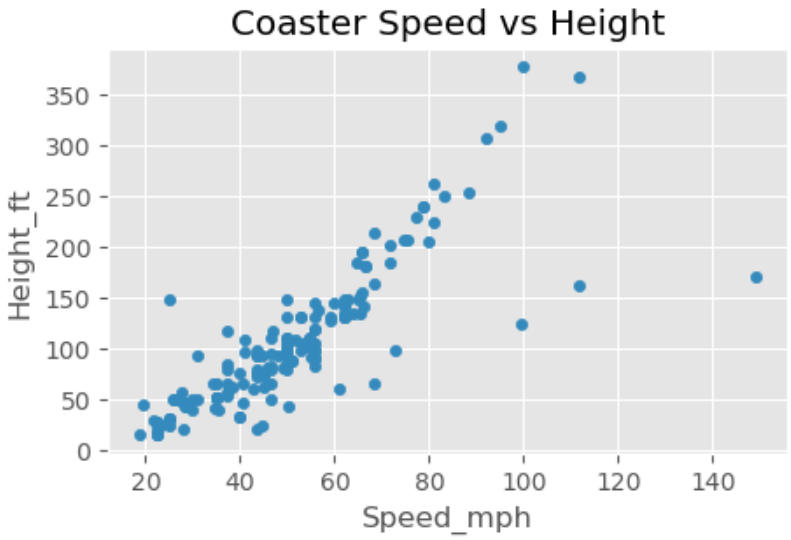
Feature In Relationships

In [279]:

```
#Scatterplot
#Heatmap Correlation
#Pairplot
#Groupby Comparison
```

In [280]:

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

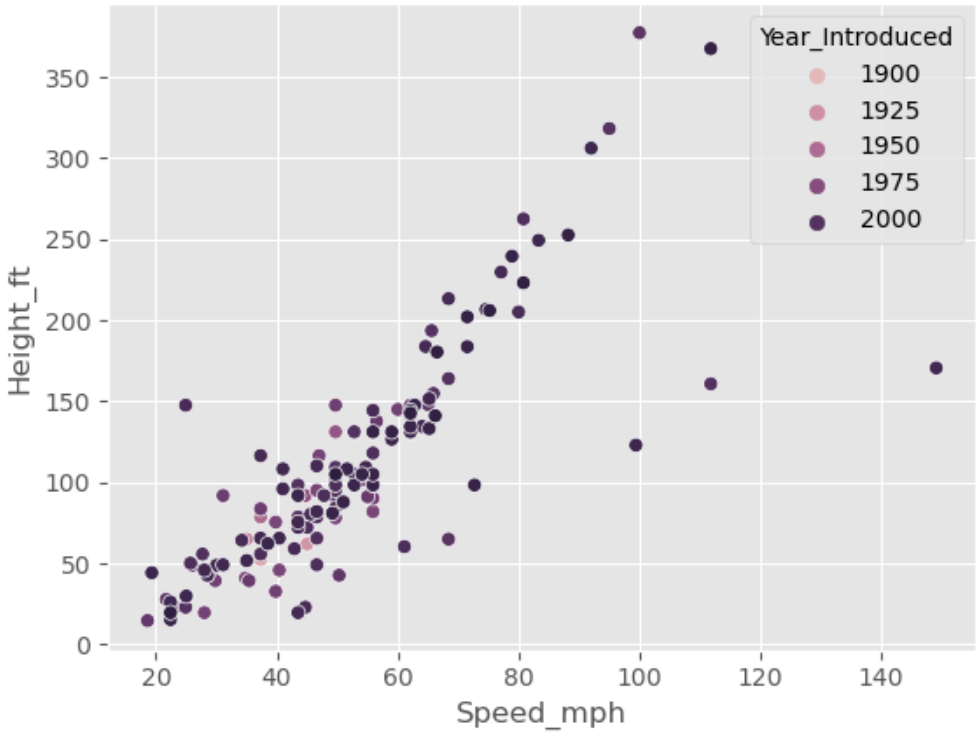


In [281]:

```
sns.scatterplot(x = 'Speed_mph',y ='Height_ft',hue = 'Year_Introduced',data = df)
```

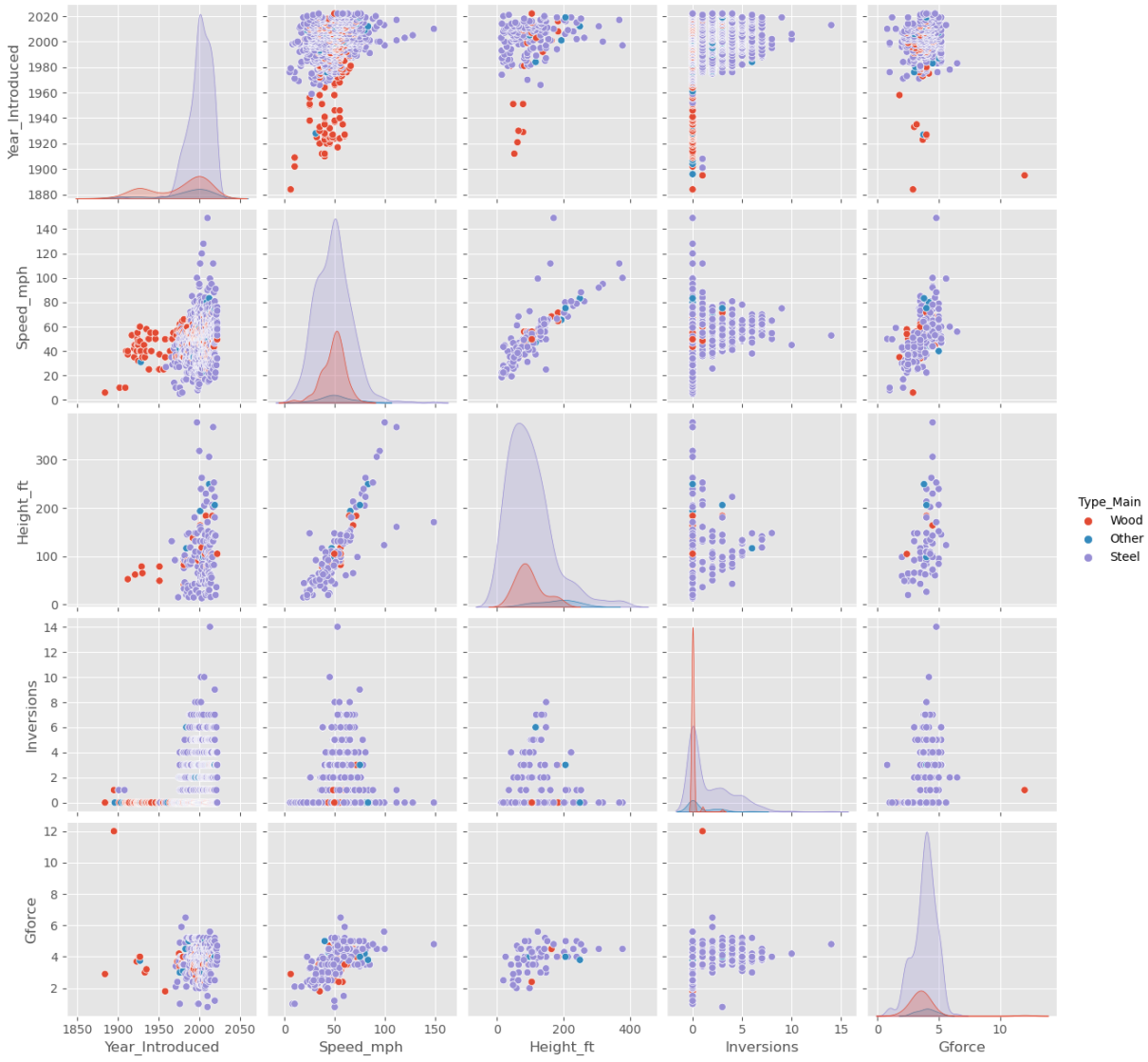
Out[281]:

```
<AxesSubplot:xlabel='Speed_mph', ylabel='Height_ft'>
```



In [282]:

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



In [283]:

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

Out[283]:

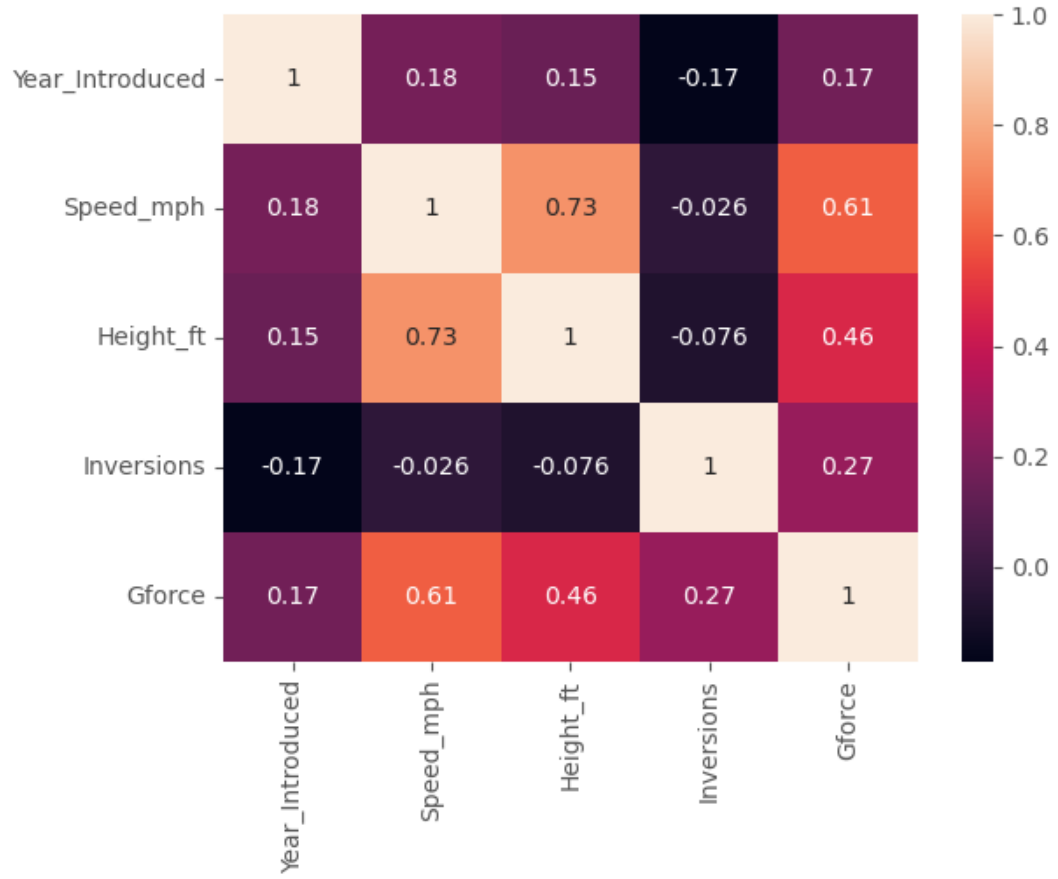
	Year_Introduced	Speed_mph	Height_ft	Inversions	Gforce
Year_Introduced	1.000000	0.178619	0.145457	-0.172829	0.168763
Speed_mph	0.178619	1.000000	0.734499	-0.026413	0.605090
Height_ft	0.145457	0.734499	1.000000	-0.076255	0.460841
Inversions	-0.172829	-0.026413	-0.076255	1.000000	0.270942
Gforce	0.168763	0.605090	0.460841	0.270942	1.000000

In [284]:

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

Out[284]:

<AxesSubplot:>



What are the locations with the fastest roller coaster(minimum of 10)?

In [285]:

```
ax = df.query('Location != "Other"').groupby('Location')['Speed_mph'].agg(['mean','count']).query('count
                                .sort_values('mean')['mean'].plot(kind = 'barh',title = 'Average Coaster S
ax.set_xlabel('Average Coaster Speed')
plt.show()
```