Exploratory Data Analysis

Summary Stats

In [2]:
```
import pandas as pd
import numpy as np
import seaborn as sns
Angel_market = pd.read_csv("angels_market.csv")
Angel_market.head()
```

Out[2]:

	vendorID	theme	homeState	carnivals	complaints	est_energy	est_hourly_vol	LL_pa
0	1	Hot Chocolate/Warm Treats	Maine	3	9	57.291961	118	
1	2	Local Artists	Vermont	1	2	39.404898	105	
2	3	Fortune Teller	New Hampshire	5	4	47.175958	94	
3	4	Fried Dough and Pizza	Maine	8	0	58.192568	118	
4	5	craft beer	New Hampshire	7	6	56.657908	102	

Removing the data which are not required. i.e. the rows where state value is integer and it doesn't make sense.

In [4]:
```
Possible_state = ['Maine','Vermont','New Hampshire','Quebec','Connecticu
t','Massachusetts','Ontario']
```

In [5]:
```
Angel_market = Angel_market[Angel_market.homeState.isin(Possible_state)]
```

In [6]:
```
Angel_market.columns
```

Out[6]:
```
Index(['vendorID', 'theme', 'homeState', 'carnivals', 'complaints',
       'est_energy', 'est_hourly_vol', 'LL_passholder', 'est_hourly_gro
ss'],
      dtype='object')
```

We will remove the vendor ID column as this may not be required in the data visualization and analysis.

In [7]:
```
Angel_market = Angel_market.drop('vendorID',1)
```

In [8]:
```
Angel_market.columns
```

Out[8]:
```
Index(['theme', 'homeState', 'carnivals', 'complaints', 'est_energy',
       'est_hourly_vol', 'LL_passholder', 'est_hourly_gross'],
      dtype='object')
```

Let's ensure that the dataset doesn't contain any NaN value.

In [9]:
```
Angel_market.isnull().values.any()
```

Out[9]:
```
False
```

In [10]:
```
Angel_market.head()
```

Out[10]:

	theme	homeState	carnivals	complaints	est_energy	est_hourly_vol	LL_passholder
0	Hot Chocolate/Warm Treats	Maine	3	9	57.291961	118	0
1	Local Artists	Vermont	1	2	39.404898	105	1
2	Fortune Teller	New Hampshire	5	4	47.175958	94	0
3	Fried Dough and Pizza	Maine	8	0	58.192568	118	0
4	craft beer	New Hampshire	7	6	56.657908	102	0

From the first 5 rows, from the angles_market dataset we can say that it consists of 8 variables.

In [11]:
```
Angel_market.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 692 entries, 0 to 691
Data columns (total 8 columns):
theme                692 non-null object
homeState            692 non-null object
carnivals            692 non-null int64
complaints           692 non-null int64
est_energy           692 non-null float64
est_hourly_vol       692 non-null int64
LL_passholder        692 non-null int64
est_hourly_gross     692 non-null float64
dtypes: float64(2), int64(4), object(2)
memory usage: 48.7+ KB
```

The dataframe has 692 rows and 8 columns. Out of the 8 variables, 6 variables are numerical(4 integers and 2 float) and remaining variables are categorical. There is no null value in the dataset.

In [12]:
```
Angel_market.describe()
```

Out[12]:

	carnivals	complaints	est_energy	est_hourly_vol	LL_passholder	est_hourly_gross
count	692.000000	692.000000	692.000000	692.000000	692.000000	692.000000
mean	5.114162	5.228324	47.975867	111.361272	0.195087	218.970159
std	2.188637	4.936342	13.544075	11.305755	0.396554	35.089505
min	0.000000	0.000000	3.069903	82.000000	0.000000	81.290000
25%	4.000000	0.000000	39.920280	104.000000	0.000000	194.295000
50%	5.000000	4.000000	48.075756	111.000000	0.000000	217.735000
75%	7.000000	9.000000	57.372752	119.000000	0.000000	242.735000
max	13.000000	20.000000	91.567936	147.000000	1.000000	322.570000

The above table provides us a quick summary of the data frame. The describe() gives summary statistics of all numeric variables. The result usually helps analyst to identify how data is distributed and which column will be helpful for the model. Each variables are measured on different unit.
The range of number of carnivals attended by the vendor listed in the data frame is 0-13 and on an average 5 carnivals were attended by the vendors from the dataset. Also, Average hourly revenue generated by the vendors ranges from $81 to 322$.

In [13]:
```
Angel_market.groupby(['LL_passholder']).agg({
'LL_passholder': ['mean','count'],
}).round(2)
```

Out[13]:

	LL_passholder	
	mean	count
LL_passholder		
0	0	557
1	1	135

In the dataset containing information about 692 vendors, 557 vendors don't hold lobster land pass whereas 135 vendors are Lobster Land passholders.

In [14]:
```
Angel_market.groupby(['LL_passholder']).agg({
'complaints': ['mean','count'],
'carnivals': ['mean'],
'est_energy' : ['mean'],
'est_hourly_gross' : ['mean'],
'est_hourly_vol' : ['mean']
}).round(2)
```

Out[14]:

	complaints		carnivals	est_energy	est_hourly_gross	est_hourly_vol
	mean	count	mean	mean	mean	mean
LL_passholder						
0	5.23	557	5.13	48.12	218.24	111.23
1	5.21	135	5.06	47.37	221.97	111.91

In the above table, we wanted to know if vendors who hold pass vs. who don't hold pass have any difference in terms of available variables. But it doesn't seem that way. This way we can tell that it might not matter who holds pass or not while allocating stalls at the winter land festival.

In [28]:
```
pd.pivot_table(Angel_market,'complaints',['theme'],aggfunc=np.mean)
```

Out[28]:

	complaints
theme	
Canadian Snacks	4.972973
DIY Ice Sculpture	4.631579
Fortune Teller	7.444444
Fried Dough and Pizza	4.800000
Games Of Chance	4.317647
Homemade Holiday Gifts	6.269231
Hot Chocolate/Warm Treats	5.265487
Local Artists	5.270270
Local Politician	6.200000
Maine Tourism Promotion	3.800000
Specialty Ice Cream	4.500000
Steaming Hot Cocktails	5.547619
Video Game/eSports	5.086957
craft beer	6.684211

From the above pivot table, we can say that there are lot of complaints from vendors who had theme such as Fortune teller, craft beer, Homemade holiday gifts and local politicians. Specialty ice cream , games of chance, canadian snacks have received the least complaints.
Regarding "Homemade holiday gifts" theme it might be possible that people might not have liked the quality of the gifts. Lobster land might want to choose theme like - Maine Tourism promotions, DIY, Ice creams etc.

In [23]:
```
pd.pivot_table(Angel_market,'est_hourly_vol',['theme'],aggfunc=np.mean)
```

Out[23]:

theme	est_hourly_vol
Canadian Snacks	111.729730
DIY Ice Sculpture	114.842105
Fortune Teller	112.444444
Fried Dough and Pizza	110.666667
Games Of Chance	111.952941
Homemade Holiday Gifts	111.423077
Hot Chocolate/Warm Treats	111.831858
Local Artists	112.216216
Local Politician	111.100000
Maine Tourism Promotion	108.066667
Specialty Ice Cream	109.766667
Steaming Hot Cocktails	110.071429
Video Game/eSports	109.260870
craft beer	110.210526

The estimated number of people visiting the different themed booths in an hour is approximately the same. There is no notable differences between the different themes' number of hourly visitors. It can however be seen that DIY Ice Scuplture, local artists and fortune tellers are the most visited.

In [24]:
```
pd.pivot_table(Angel_market,'est_hourly_gross',['theme'],aggfunc=np.mean
)
```

Out[24]:

theme	est_hourly_gross
Canadian Snacks	221.436892
DIY Ice Sculpture	222.981053
Fortune Teller	207.072222
Fried Dough and Pizza	219.167333
Games Of Chance	222.085176
Homemade Holiday Gifts	215.885385
Hot Chocolate/Warm Treats	214.720354
Local Artists	224.376216
Local Politician	222.541000
Maine Tourism Promotion	215.484000
Specialty Ice Cream	217.727333
Steaming Hot Cocktails	218.952381
Video Game/eSports	217.996522
craft beer	221.432105

It can be seen that most revenue is expected to be generated by DIY Ice Scuplture, local artists, Games of Chance, local politican, craft beer and canadian snacks.

In [39]:
```
AM_pivot = pd.pivot_table(Angel_market,'complaints',index = ['theme'],co
lumns = ['homeState'],aggfunc=np.mean)
```

In [40]:
```
AM_pivot.head(15)
```

Out[40]:

homeState	Connecticut	Maine	Massachusetts	New Hampshire	Ontario	Quebec	Vermont
theme							
Canadian Snacks	4.857143	4.675000	5.000000	4.111111	8.0	6.800000	5.555556
DIY Ice Sculpture	NaN	6.111111	NaN	3.400000	NaN	0.500000	5.000000
Fortune Teller	14.000000	6.250000	15.000000	4.000000	NaN	3.000000	6.000000
Fried Dough and Pizza	5.000000	3.666667	NaN	5.300000	0.0	7.200000	9.166667
Games Of Chance	2.444444	3.910714	6.500000	5.500000	15.0	5.428571	4.000000
Homemade Holiday Gifts	10.250000	6.064516	7.500000	3.333333	8.0	9.666667	4.076923
Hot Chocolate/Warm Treats	5.700000	4.633803	13.333333	6.545455	3.2	6.857143	5.500000
Local Artists	0.000000	5.581395	2.000000	4.625000	5.0	5.800000	7.600000
Local Politician	0.000000	5.625000	17.000000	NaN	NaN	NaN	NaN
Maine Tourism Promotion	NaN	3.800000	NaN	NaN	NaN	NaN	NaN
Specialty Ice Cream	7.166667	4.777778	0.000000	1.000000	NaN	1.333333	NaN
Steaming Hot Cocktails	NaN	5.277778	6.500000	3.375000	0.0	4.250000	9.000000
Video Game/eSports	0.000000	5.800000	17.000000	0.000000	NaN	6.000000	1.750000
craft beer	0.000000	7.230769	NaN	4.000000	NaN	12.000000	9.000000

In [19]:
```
Angel_market.groupby(['homeState']).agg({
'homeState': ['count'],
}).round(2)
```

Out[19]:

	homeState
	count
homeState	
Connecticut	48
Maine	417
Massachusetts	27
New Hampshire	70
Ontario	16
Quebec	54
Vermont	60

The highest number of vendors i.e. 417 vendors belong to Maine, 70 from New Hampshire, 60 from Vermont, 48 from Connecticut and 27 from Massachusetts. Only 10% of the total vendors belong to Canada - 54 from Quebec and 16 from Ontario.

In [38]:
```
AM_pivot1 = pd.pivot_table(Angel_market,'complaints',index = ['theme'],c
olumns = ['homeState'],aggfunc=np.mean,margins= True)
```

In [42]:
```
AM_pivot1.head(15)
```

Out[42]:

homeState	Connecticut	Maine	Massachusetts	New Hampshire	Ontario	Quebec	Vermont
theme							
Canadian Snacks	4.857143	4.675000	5.000000	4.111111	8.0000	6.800000	5.555556
DIY Ice Sculpture	NaN	6.111111	NaN	3.400000	NaN	0.500000	5.000000
Fortune Teller	14.000000	6.250000	15.000000	4.000000	NaN	3.000000	6.000000
Fried Dough and Pizza	5.000000	3.666667	NaN	5.300000	0.0000	7.200000	9.166667
Games Of Chance	2.444444	3.910714	6.500000	5.500000	15.0000	5.428571	4.000000
Homemade Holiday Gifts	10.250000	6.064516	7.500000	3.333333	8.0000	9.666667	4.076923
Hot Chocolate/Warm Treats	5.700000	4.633803	13.333333	6.545455	3.2000	6.857143	5.500000
Local Artists	0.000000	5.581395	2.000000	4.625000	5.0000	5.800000	7.600000
Local Politician	0.000000	5.625000	17.000000	NaN	NaN	NaN	NaN
Maine Tourism Promotion	NaN	3.800000	NaN	NaN	NaN	NaN	NaN
Specialty Ice Cream	7.166667	4.777778	0.000000	1.000000	NaN	1.333333	NaN
Steaming Hot Cocktails	NaN	5.277778	6.500000	3.375000	0.0000	4.250000	9.000000
Video Game/eSports	0.000000	5.800000	17.000000	0.000000	NaN	6.000000	1.750000
craft beer	0.000000	7.230769	NaN	4.000000	NaN	12.000000	9.000000
All	5.562500	4.940048	7.296296	4.485714	6.1875	5.962963	5.983333

Summary stats has been used to explore the dataset angels_market, which provides information about 692 vendors who will be a part of the Winter Wonderland at Lobster Land. Almost 75% of the vendors have attended at least 4 carnivals before, hence comparison with previous performances will help improve. On an average, each vendor's booth will be visited by 111 people in an hour, which indicates a good footfall. The hourly revenue for the vendors is expected to be between 81 dollars to 322 dollars, with an average revenue of 218 dollars. It is interesting to see that whether the vendor is a passholder or not has no impact on the variables.
For any event, it is important to understand which stalls have garnered the most complaints in the past. The pivot table shows that fortune teller, craft beer, homemade holiday gifts, local politician and steaming hot cocktails received the most complaints. Maine Tourism promotion had the least complaints. However, it could be so owing to the fact that it has low number of hourly visitors as compared to other booths. The park must focus on stalls wherein people can participate and interact such as video games, local artists, DIY ice sculptures, games of chance, as these show most expected hourly revenue and less complaints too.

It is interesting to see that even though 60% of the vendors (417 vendors) are from Maine, very less complaints have been received from these vendors as compared to those belonging to other states. The number of vendors belonging to Ontario, Massachusetts and Quebec is very low, but have received a very high number of complaints. The reason behind this could be the reduced quality of foods and drinks owing to transportation, customers didn't like Canadian craft beer or dissatisfaction by fortune tellers. Hence, the park must focus on vendors belonging to Maine (convenient too), improve the quality of food and drinks, focus on interactive booths.