California Grants Dataset EDA

EDA Using Python

- understand data
  - many columns are not very useful
  - some that may be useful are self reported, having no consistency in formatting between rows (award period, estimated amounts.)
- clean data
- analyze variables

In [6]:
```
# import necessary packages
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [7]:
```
grants_raw = pd.read_csv("ca_grants.csv")
```

In [8]:
```
grants_raw.head()
```

Out[8]:

	PortalID	GrantID	Status	LastUpdated	ChangeNotes	AgencyDept	Title	Type	LOI	
0	6481	DOJ-PROP56-2022-23-1	active	2022-07-18 17:28:34	Updated eligibility, suggested activities, and...	Department of Justice (Office of the Attorney ...	Tobacco Grant Program FY 2022-23 Request for P...	Grant	No	
1	11966	NaN	active	2022-07-15 22:20:56	Application open date: July 15, 2022. Updated...	Department of Health Care Access and Information	California State Loan Repayment Program (SLRP)	Grant	No	
2	11960	NaN	active	2022-07-14 22:35:54	NaN	CA Arts Council	Administering Organization – Individual Artist...	Grant	No	Di C
3	11957	NaN	active	2022-07-14 22:15:54	NaN	CA Arts Council	Administering Organization – Arts Administrato...	Grant	No	Di C
4	11912	NaN	active	2022-07-14 17:13:34	NaN	Department of Pesticide Regulation	Department of Pesticide Regulation 2023 Allian...	Grant	No	Di C

5 rows × 36 columns

In [9]:
```
print(grants_raw.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 725 entries, 0 to 724
Data columns (total 36 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   PortalID             725 non-null    int64 
 1   GrantID              132 non-null    object
 2   Status               725 non-null    object
 3   LastUpdated          725 non-null    object
 4   ChangeNotes          419 non-null    object
 5   AgencyDept           725 non-null    object
 6   Title                725 non-null    object
 7   Type                 725 non-null    object
 8   LOI                  724 non-null    object
 9   Categories           725 non-null    object
 10  CategorySuggestion   94 non-null     object
 11  Purpose              724 non-null    object
 12  Description          725 non-null    object
 13  ApplicantType        721 non-null    object
 14  ApplicantTypeNotes   604 non-null    object
 15  Geography            557 non-null    object
 16  FundingSource        721 non-null    object
 17  FundingSourceNotes   523 non-null    object
 18  MatchingFunds        725 non-null    object
 19  MatchingFundsNotes   279 non-null    object
 20  EstAvailFunds        713 non-null    object
 21  EstAwards            725 non-null    object
 22  EstAmounts           725 non-null    object
 23  FundingMethod        719 non-null    object
 24  FundingMethodNotes   426 non-null    object
 25  OpenDate             720 non-null    object
 26  ApplicationDeadline  715 non-null    object
 27  AwardPeriod          712 non-null    object
 28  ExpAwardDate         713 non-null    object
 29  ElecSubmission       609 non-null    object
 30  GrantURL             725 non-null    object
 31  AgencyURL            724 non-null    object
 32  AgencySubscribeURL   400 non-null    object
 33  GrantEventsURL       275 non-null    object
 34  ContactInfo          725 non-null    object
 35  AwardStats           610 non-null    object
dtypes: int64(1), object(35)
memory usage: 204.0+ KB
None
```

In [10]:
```
# Change ID of the grant to a categorical variable
grants_raw['PortalID'] = grants_raw['PortalID'].astype('object')
```

Removal of unecessary columns

- columns with excessive missing values
- redundant information
- information that cannot realistically be useful or analyzed

In [11]:
```
grants = grants_raw.drop(grants_raw.columns[[1,4,10,14,15,17,19,24,27,29,31,32,33,34,3
grants.head()
```

Out[11]:

	PortalID	Status	LastUpdated	AgencyDept	Title	Type	LOI	Categories	Purpos
0	6481	active	2022-07-18 17:28:34	Department of Justice (Office of the Attorney ...	Tobacco Grant Program FY 2022-23 Request for P...	Grant	No	Education; Law, Justice, and Legal Services	The purpos of this gra offere through the
1	11966	active	2022-07-15 22:20:56	Department of Health Care Access and Information	California State Loan Repayment Program (SLRP)	Grant	No	Health & Human Services	The Californ State Loa Repayme Program (S
2	11960	active	2022-07-14 22:35:54	CA Arts Council	Administering Organization – Individual Artist...	Grant	No	Disadvantaged Communities; Libraries and Arts	Th Administerir Organization Individual Ar
3	11957	active	2022-07-14 22:15:54	CA Arts Council	Administering Organization – Arts Administrato...	Grant	No	Disadvantaged Communities; Education; Employme...	The Ar Administrato Pipelir Fellowship pr
4	11912	active	2022-07-14 17:13:34	Department of Pesticide Regulation	Department of Pesticide Regulation 2023 Allian...	Grant	No	Agriculture; Disadvantaged Communities; Educat...	To promo safer, mo sustainab pest manage

5 rows × 21 columns

In [12]:
```
# Columns we are left with
print(grants.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 725 entries, 0 to 724
Data columns (total 21 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   PortalID             725 non-null    object
 1   Status               725 non-null    object
 2   LastUpdated          725 non-null    object
 3   AgencyDept           725 non-null    object
 4   Title                725 non-null    object
 5   Type                 725 non-null    object
 6   LOI                  724 non-null    object
 7   Categories           725 non-null    object
 8   Purpose              724 non-null    object
 9   Description          725 non-null    object
 10  ApplicantType        721 non-null    object
 11  FundingSource        721 non-null    object
 12  MatchingFunds        725 non-null    object
 13  EstAvailFunds        713 non-null    object
 14  EstAwards            725 non-null    object
 15  EstAmounts           725 non-null    object
 16  FundingMethod        719 non-null    object
 17  OpenDate             720 non-null    object
 18  ApplicationDeadline  715 non-null    object
 19  ExpAwardDate         713 non-null    object
 20  GrantURL             725 non-null    object
dtypes: object(21)
memory usage: 119.1+ KB
None
```

In [13]:
```
grants.describe(include=["object"]) #no duplicate rows
```

Out[13]:

	PortalID	Status	LastUpdated	AgencyDept	Title	Type	LOI	Categories	Purpose	D
count	725	725	725	725	725	725	724	725	724	
unique	725	3	725	61	659	3	2	196	628	
top	6481	closed	2022-07-18 17:28:34	Governor's Office of Emergency Services	Vertebrate Pest Control Research Program	Grant	No	Environment & Water	The purpose of the Cannabis Tax Fund Grant Pro...	P
freq	1	572	1	48	4	657	576	94	7	

4 rows × 21 columns

Converting the columns 'EstAwards', 'EstAmounts', and 'EstAvailFunds' into numeric variables. Unique values reveal that the entries for these two columns are formatted consistently. As many entries contain a range of values these columns were each split into 2, relecting their maximum and minimum values. Undeclared entries were replaces with a missing value (NaN).

In [14]:
```
grants['EstAwards'].unique()
```

Out[14]:
```
array(['Dependant on number of submissions received, application process, etc.',
       'Exactly 1', 'Between 1 and 30', 'Between 1 and 63',
       'Between 1 and 6', 'Exactly 21', 'Exactly 10', 'Between 1 and 2',
       'Between 50 and 60', 'Between 4 and 7', 'Exactly 13', 'Exactly 2',
       'Exactly 5', 'Exactly 14', 'Exactly 20', 'Exactly 4', 'Exactly 3',
       'Between 35 and 35', 'Between 25 and 40', 'Between 2 and 10',
       'Between 500 and 1000', 'Between 30 and 35', 'Between 0 and 0',
       'Between 0 and 77', 'Exactly 7', 'Exactly 50', 'Between 10 and 20',
       'Between 0 and 100', 'Exactly 40', 'Between 0 and 20',
       'Between 1 and 8', 'Exactly 70', 'Exactly 6', 'Between 8 and 12',
       'Exactly 100', 'Between 0 and 62732', 'Between 0 and 270',
       'Between 0 and 675', 'Between 15 and 30', 'Exactly 225',
       'Between 2 and 4', 'Between 0 and 58', 'Between 1 and 17',
       'Between 20 and 30', 'Between 1 and 18',
       'Between 100000 and 500000', 'Between 1 and 10',
       'Between 60 and 79', 'Exactly 16', 'Between 1 and 12',
       'Between 65 and 85', 'Between 40 and 50', 'Between 16 and 30',
       'Between 1 and 5', 'Exactly 9', 'Between 1 and 4', 'Exactly 90',
       'Between 5 and 12', 'Exactly 8', 'Between 5 and 7',
       'Between 10 and 12', 'Between 3 and 5', 'Between 1 and 15',
       'Between 1 and 3', 'Between 0 and 3', 'Exactly 12', 'Exactly 432'],
      dtype=object)
```

In [15]:
```
awards = grants['EstAwards']
maxaward = []
minaward = []
for i in (range(len(awards))):
    if awards[i][0] == 'E':
        maxaward.append(int(''.join(filter(str.isdigit, awards[i]))))
        minaward.append(int(''.join(filter(str.isdigit, awards[i]))))
    elif awards[i][0] == 'B':
        maxaward.append(int(''.join(filter(str.isdigit, awards[i].rpartition('a')[2]))))
        minaward.append(int(''.join(filter(str.isdigit, awards[i].rpartition('a')[0]))))
    else:
        maxaward.append(float('nan'))
        minaward.append(float('nan'))


amounts = grants['EstAmounts']
maxamnt = []
minamnt = []
for i in (range(len(amounts))):
    if amounts[i][0] == 'E':
        maxamnt.append(int(''.join(filter(str.isdigit, amounts[i]))))
        minamnt.append(int(''.join(filter(str.isdigit, amounts[i]))))
    elif amounts[i][0] == 'B':
        maxamnt.append(int(''.join(filter(str.isdigit, amounts[i].rpartition('a')[2]))))
        minamnt.append(int(''.join(filter(str.isdigit, amounts[i].rpartition('a')[0]))))
    else:
        maxamnt.append(float('nan'))
        minamnt.append(float('nan'))
```

In [16]:
```
grants['MaxAwards'] = maxaward
grants['MinAwards'] = minaward
grants = grants.drop('EstAwards', axis = 1)

grants['MaxAmounts'] = maxamnt
grants['MinAmounts'] = minamnt
grants = grants.drop('EstAmounts', axis = 1)
```

In [17]:
```
availfunds = []
for i in (range(len(grants['EstAvailFunds']))):
    if type(grants['EstAvailFunds'][i]) != str:
        availfunds.append(float('nan'))
    else:
        availfunds.append(int(''.join(filter(str.isdigit, grants['EstAvailFunds'][i]))
```

In [18]:
```
grants['EstAvailFunds'] = availfunds
grants.head()
```

Out[18]:

	PortalID	Status	LastUpdated	AgencyDept	Title	Type	LOI	Categories	Purpos
0	6481	active	2022-07-18 17:28:34	Department of Justice (Office of the Attorney ...	Tobacco Grant Program FY 2022-23 Request for P...	Grant	No	Education; Law, Justice, and Legal Services	The purpos of this gra offere through the
1	11966	active	2022-07-15 22:20:56	Department of Health Care Access and Information	California State Loan Repayment Program (SLRP)	Grant	No	Health & Human Services	The Californ State Loa Repayme Program (S
2	11960	active	2022-07-14 22:35:54	CA Arts Council	Administering Organization – Individual Artist...	Grant	No	Disadvantaged Communities; Libraries and Arts	Th Administerir Organization Individual Ar
3	11957	active	2022-07-14 22:15:54	CA Arts Council	Administering Organization – Arts Administrato...	Grant	No	Disadvantaged Communities; Education; Employme...	The Ar Administrato Pipelir Fellowship pr
4	11912	active	2022-07-14 17:13:34	Department of Pesticide Regulation	Department of Pesticide Regulation 2023 Allian...	Grant	No	Agriculture; Disadvantaged Communities; Educat...	To promo safer, mo sustainab pest manage

5 rows × 23 columns

In [19]:
```
print(grants.info()) # Our new columns are left with mostly missing values as a majori
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 725 entries, 0 to 724
Data columns (total 23 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   PortalID             725 non-null    object
 1   Status               725 non-null    object
 2   LastUpdated          725 non-null    object
 3   AgencyDept           725 non-null    object
 4   Title                725 non-null    object
 5   Type                 725 non-null    object
 6   LOI                  724 non-null    object
 7   Categories           725 non-null    object
 8   Purpose              724 non-null    object
 9   Description          725 non-null    object
 10  ApplicantType        721 non-null    object
 11  FundingSource        721 non-null    object
 12  MatchingFunds        725 non-null    object
 13  EstAvailFunds        713 non-null    float64
 14  FundingMethod        719 non-null    object
 15  OpenDate             720 non-null    object
 16  ApplicationDeadline  715 non-null    object
 17  ExpAwardDate         713 non-null    object
 18  GrantURL             725 non-null    object
 19  MaxAwards            148 non-null    float64
 20  MinAwards            148 non-null    float64
 21  MaxAmounts           231 non-null    float64
 22  MinAmounts           231 non-null    float64
dtypes: float64(5), object(18)
memory usage: 130.4+ KB
None
```

In [20]:
```
grants.describe(include=["object"])
```

Out[20]:

	PortalID	Status	LastUpdated	AgencyDept	Title	Type	LOI	Categories	Purpose	D
count	725	725	725	725	725	725	724	725	724	
unique	725	3	725	61	659	3	2	196	628	
top	6481	closed	2022-07-18 17:28:34	Governor's Office of Emergency Services	Vertebrate Pest Control Research Program	Grant	No	Environment & Water	The purpose of the Cannabis Tax Fund Grant Pro...	P
freq	1	572	1	48	4	657	576	94	7	

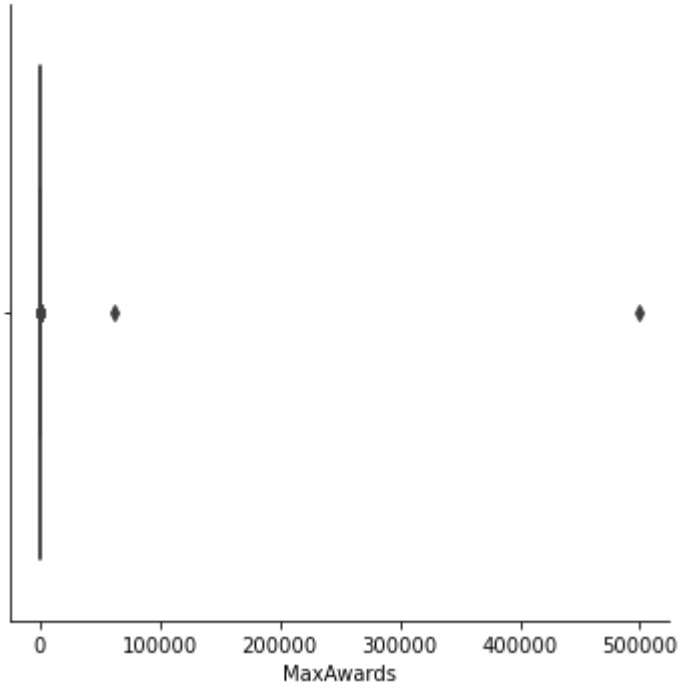
In [21]:
```
grants.describe()
```

Out[21]:

	EstAvailFunds	MaxAwards	MinAwards	MaxAmounts	MinAmounts
count	7.130000e+02	148.000000	148.000000	2.310000e+02	2.310000e+02
mean	6.380992e+07	3834.182432	693.114865	4.843356e+07	3.098185e+05
std	3.500748e+08	41384.329800	8218.712299	4.652575e+08	1.404581e+06
min	1.000000e+00	0.000000	0.000000	1.380000e+02	0.000000e+00
25%	1.170000e+06	2.000000	1.000000	1.000000e+05	0.000000e+00
50%	5.000000e+06	7.000000	2.000000	3.500000e+05	5.000000e+03
75%	2.000000e+07	20.000000	12.250000	1.500000e+06	5.000000e+04
max	5.000000e+09	500000.000000	100000.000000	5.000000e+09	1.500000e+07

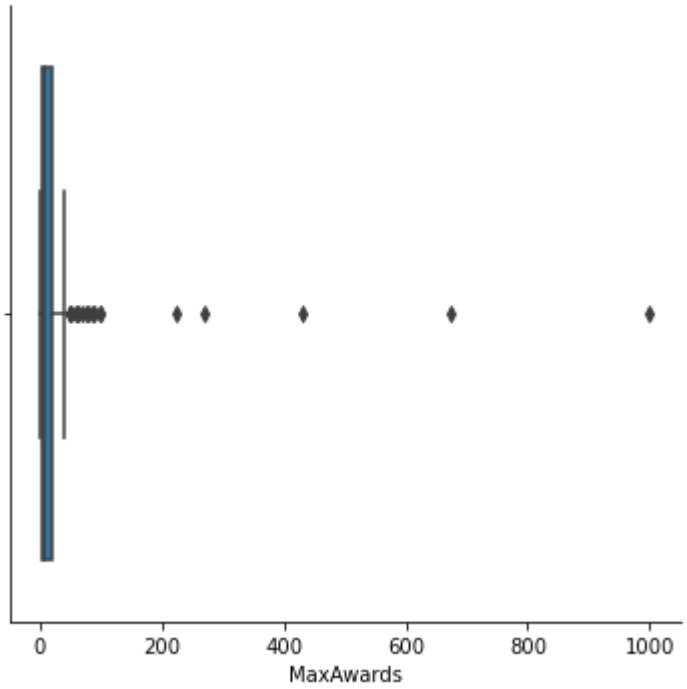
Exploring the Variables

In [23]:
```
sns.catplot(x = 'MaxAwards', kind = 'box', data = grants)
grants2 = grants[grants["MaxAwards"] < 50000] # remove the excessively large outliars
```



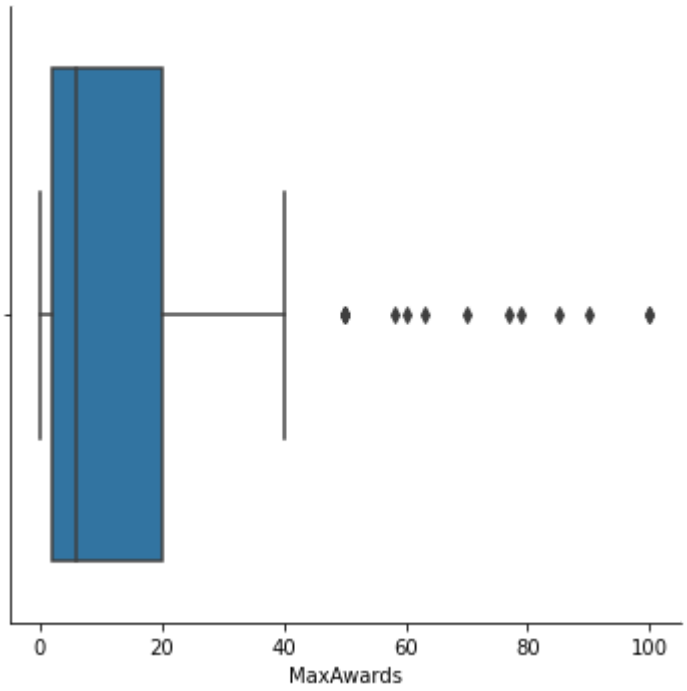
In [43]:
```
sns.catplot(x = 'MaxAwards', kind = 'box', data = grants2) # still many outliars to pc
```

Out[43]:
```
<seaborn.axisgrid.FacetGrid at 0x2610c63ed00>
```



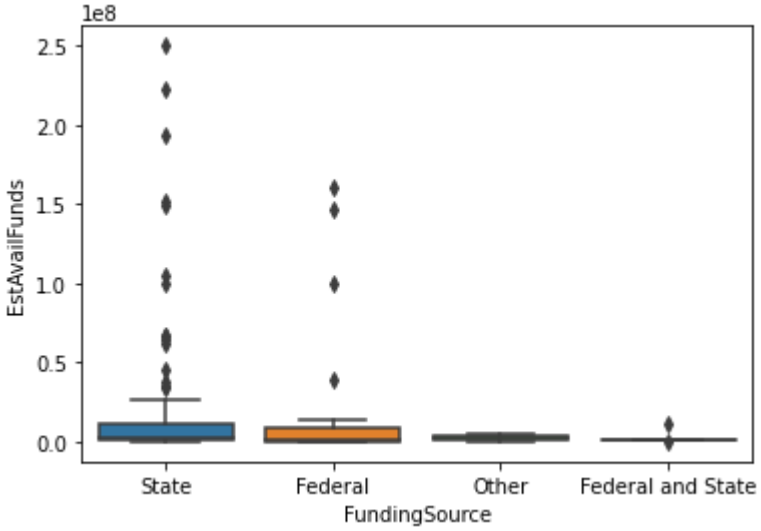
In [44]:
```
grants3 = grants[grants["MaxAwards"] < 200] # further subset our data
sns.catplot(x = 'MaxAwards', kind = 'box', data = grants3)
```

Out[44]:
```
<seaborn.axisgrid.FacetGrid at 0x26100063fd0>
```



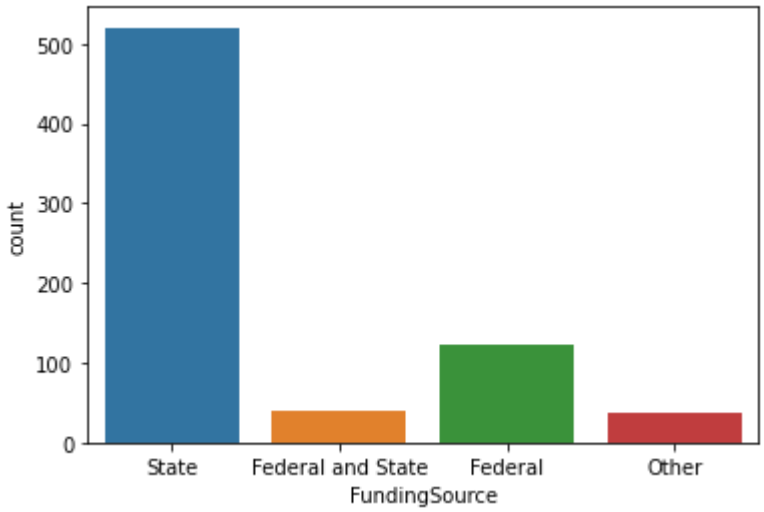
In [26]:
```
# Potential relationship: Funding Source and Maximum Awards?
sns.boxplot(x = 'FundingSource', y = 'EstAvailFunds', data = grants2) #bulk of outlier
```

Out[26]:
```
<AxesSubplot:xlabel='FundingSource', ylabel='EstAvailFunds'>
```



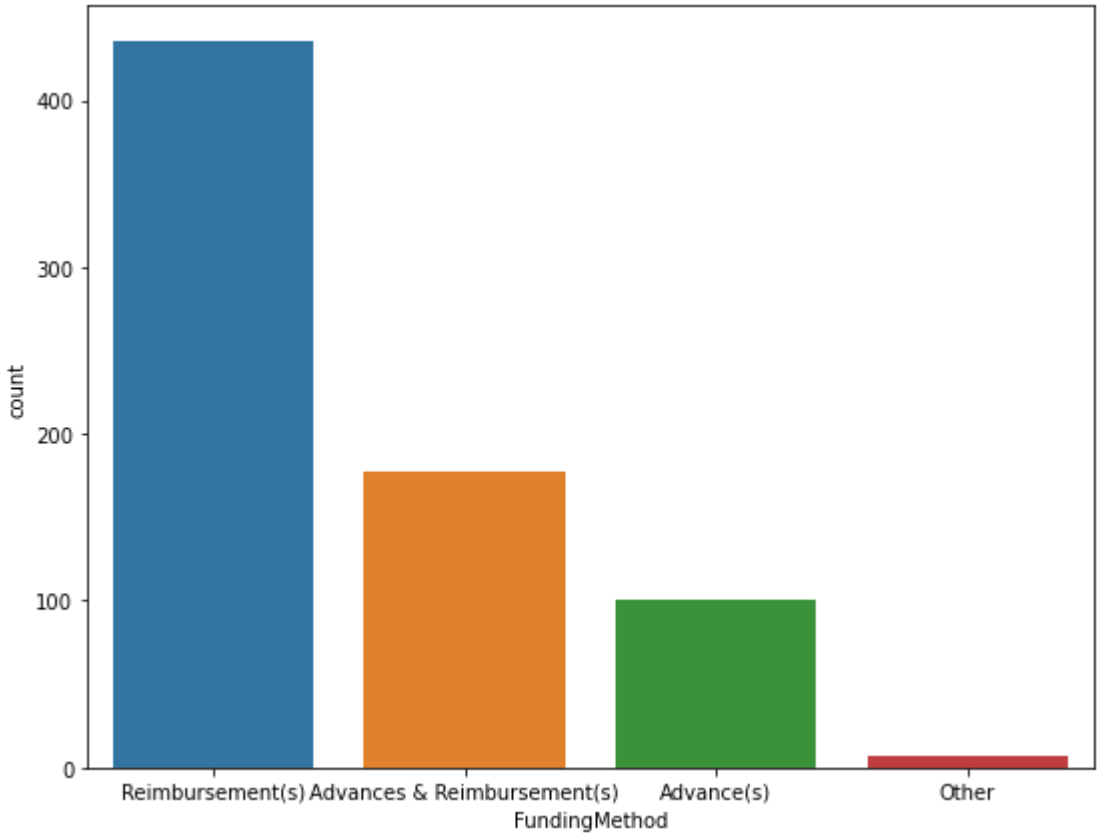
In [27]:
```
sns.countplot(x = 'FundingSource', data = grants) #to be expected as we are dealing wi
```

Out[27]:
```
<AxesSubplot:xlabel='FundingSource', ylabel='count'>
```



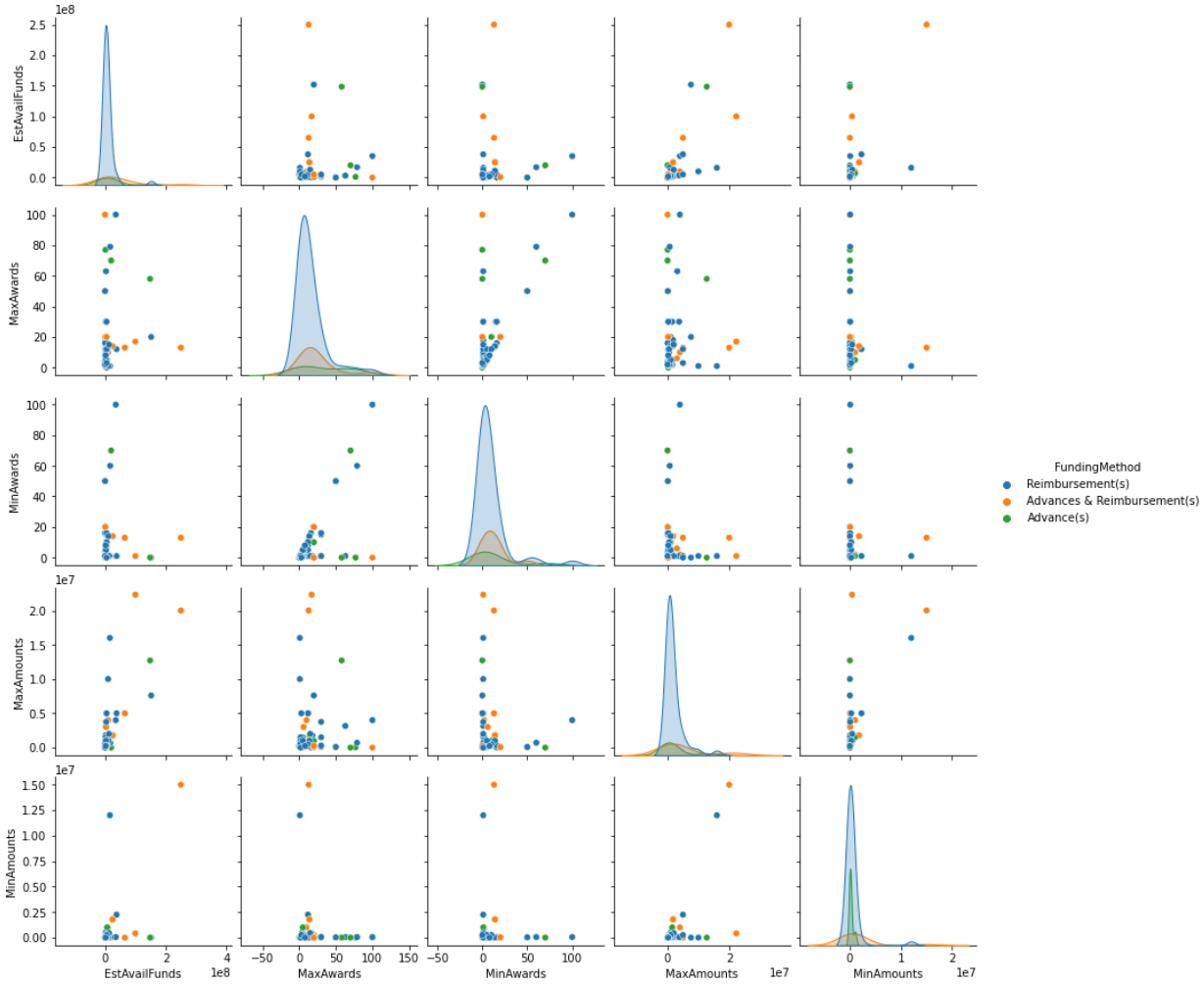
In [28]:
```
# Another potentially interesting variable to consider: Funding Method
fig, ax = plt.subplots()
fig.set_size_inches(9,7)
sns.countplot(x = 'FundingMethod', data = grants, ax = ax)
```

Out[28]:
```
<AxesSubplot:xlabel='FundingMethod', ylabel='count'>
```



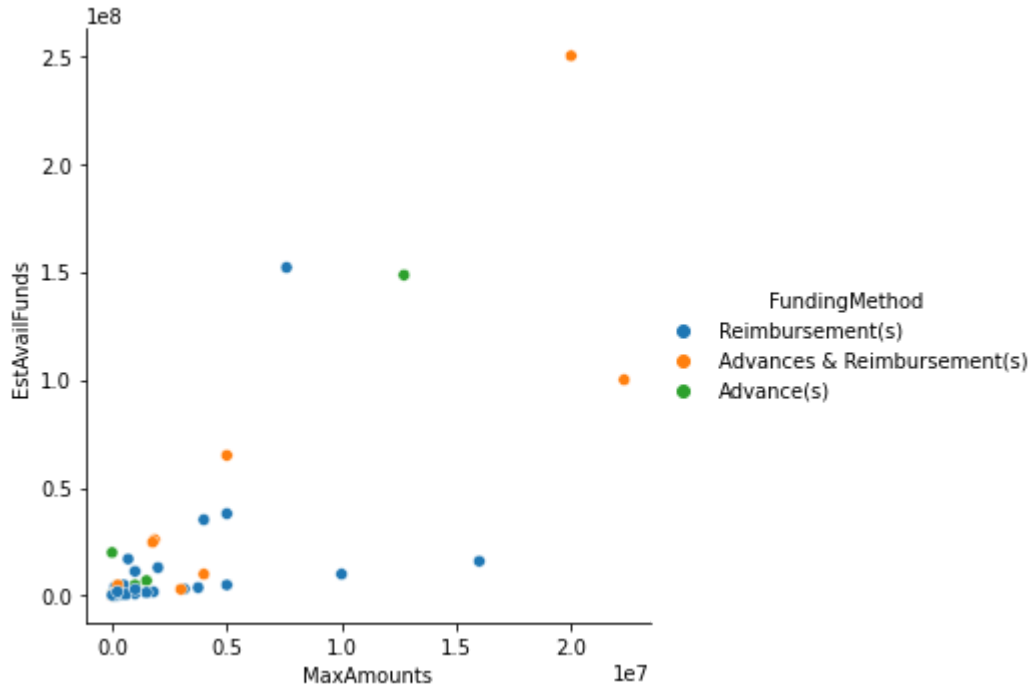
In [42]:
```
# Looking further into funding method
grants4 = grants3.dropna(axis=0, how='any', thresh=None, subset=None, inplace=False) #
sns.pairplot(data = grants4.drop('PortalID', axis = 1), hue = 'FundingMethod')
```

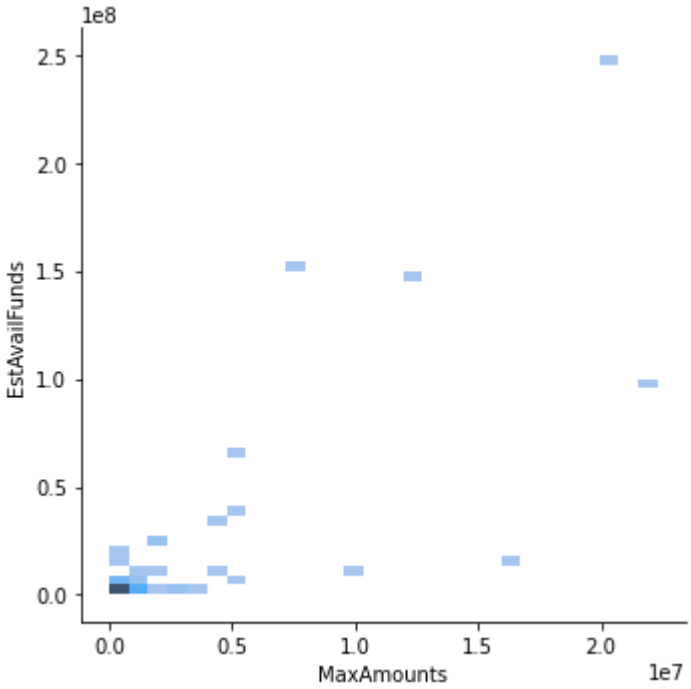
Out[42]:
```
<seaborn.axisgrid.PairGrid at 0x2610a6a2df0>
```



In [41]:
```
# Most interesting scatter: Maximum Amount vs Estimated Available Funds?
sns.relplot(x = 'MaxAmounts', y = 'EstAvailFunds', hue = 'FundingMethod', data = grant
sns.displot(data = grants4, x = 'MaxAmounts', y = 'EstAvailFunds')
```

Out[41]:
```
<seaborn.axisgrid.FacetGrid at 0x2610a5d1d00>
```





Next Steps

- apply transformation
- linear model/analysis
- potential multivariate analysis as well?