In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(color_codes=True)
```

```
Matplotlib is building the font cache; this may take a moment.
```

In [30]:
```
survey = pd.read_csv("C:/Users/camil/Downloads/survey.csv")
survey.head()
```

Out[30]:

	Year	Industry_aggregation_NZSIOC	Industry_code_NZSIOC	Industry_name_NZSIOC	Units	Variable_code	Variable_name
0	2019	Level 1	99999	All industries	Dollars (millions)	H01	Total income
1	2019	Level 1	99999	All industries	Dollars (millions)	H04	Sales, government funding, grants and subsidies
2	2019	Level 1	99999	All industries	Dollars (millions)	H05	Interest, dividends and donations
3	2019	Level 1	99999	All industries	Dollars (millions)	H07	Non-operating income
4	2019	Level 1	99999	All industries	Dollars (millions)	H08	Total expenditure

In [26]:
```
survey.columns
```

Out[26]:
```
Index(['Year', 'Industry_aggregation_NZSIOC', 'Industry_code_NZSIOC',
       'Industry_name_NZSIOC', 'Units', 'Variable_code', 'Variable_name',
       'Variable_category', 'Value', 'Industry_code_ANZSIC06'],
      dtype='object')
```

In [28]:
```
survey.columns.values
```

Out[28]:
```
array(['Year', 'Industry_aggregation_NZSIOC', 'Industry_code_NZSIOC',
       'Industry_name_NZSIOC', 'Units', 'Variable_code', 'Variable_name',
       'Variable_category', 'Value', 'Industry_code_ANZSIC06'],
      dtype=object)
```

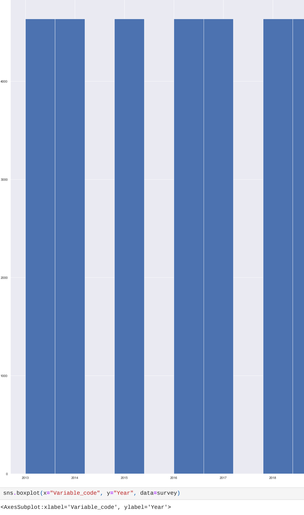
In [29]:
```
survey.describe(include='all')
```

Out[29]:

	Year	Industry_aggregation_NZSIOC	Industry_code_NZSIOC	Industry_name_NZSIOC	Units	Variable_code	Va
count	32445.000000	32445	32445	32445	32445	32445	
unique	NaN	3	139	119	3	39	
top	NaN	Level 4	CC611	Public Order, Safety and Regulatory Services	Dollars (millions)	H20	
freq	NaN	17759	252	589	25508	973	
mean	2016.000000	NaN	NaN	NaN	NaN	NaN	
std	2.000031	NaN	NaN	NaN	NaN	NaN	
min	2013.000000	NaN	NaN	NaN	NaN	NaN	
25%	2014.000000	NaN	NaN	NaN	NaN	NaN	
50%	2016.000000	NaN	NaN	NaN	NaN	NaN	
75%	2018.000000	NaN	NaN	NaN	NaN	NaN	
max	2019.000000	NaN	NaN	NaN	NaN	NaN	

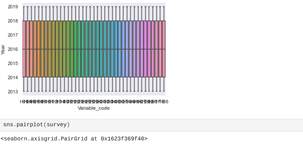
In [31]:
```
survey.hist(figsize=(20,30))
```

Out[31]:
```
array([[<AxesSubplot:title={'center':'Year'}>]], dtype=object)
```



In [36]:
```
sns.boxplot(x="Variable_code", y="Year", data=survey)
```

Out[36]:
```
<AxesSubplot:xlabel='Variable_code', ylabel='Year'>
```



In [37]:
```
sns.pairplot(survey)
```

Out[37]:
```
<seaborn.axisgrid.PairGrid at 0x1623f369f40>
```

In [43]:
```
survey= survey.rename(columns={"Industry_aggregation_NZSIOC" : "Industry_aggregation", "Industry_code
survey.head(5)
```

In [43]:

	Year	Industry_aggregation	Industry_code	Industry_name	Units	Variable_code	Variable_name	Variable_category	Valu
0	2019	Level 1	99999	All industries	Dollars (millions)	H01	Total income	Financial performance	728.23
1	2019	Level 1	99999	All industries	Dollars (millions)	H04	Sales, government funding, grants and subsidies	Financial performance	643.80
2	2019	Level 1	99999	All industries	Dollars (millions)	H05	Interest, dividends and donations	Financial performance	62.92
3	2019	Level 1	99999	All industries	Dollars (millions)	H07	Non-operating income	Financial performance	21.50
4	2019	Level 1	99999	All industries	Dollars (millions)	H08	Total expenditure	Financial performance	634.71

In [44]:
```
survey.shape
```

Out[44]:
```
(32445, 10)
```

In [46]:
```
duplicate_rows_survey =survey[survey.duplicated()]
print("Number of duplicate rows: ", duplicate_rows_survey.shape)
```

```
Number of duplicate rows:  (0, 10)
```

In [47]:
```
survey.count()
```

Out[47]:
```
Year                    32445
Industry_aggregation    32445
Industry_code           32445
Industry_name           32445
Units                   32445
Variable_code           32445
Variable_name           32445
Variable_category       32445
Value                   32445
Industry_code_ANZSIC06  32445
dtype: int64
```

In [48]:
```
survey=survey.drop_duplicates()
survey.head()
```

Out[48]:

	Year	Industry_aggregation	Industry_code	Industry_name	Units	Variable_code	Variable_name	Variable_category	Valu
0	2019	Level 1	99999	All industries	Dollars (millions)	H01	Total income	Financial performance	728.23
1	2019	Level 1	99999	All industries	Dollars (millions)	H04	Sales, government funding, grants and subsidies	Financial performance	643.80
2	2019	Level 1	99999	All industries	Dollars (millions)	H05	Interest, dividends and donations	Financial performance	62.92
3	2019	Level 1	99999	All industries	Dollars (millions)	H07	Non-operating income	Financial performance	21.50
4	2019	Level 1	99999	All industries	Dollars (millions)	H08	Total expenditure	Financial performance	634.71

In [49]:
```
survey=survey.drop_duplicates()
survey.count()
```

Out[49]:
```
Year                    32445
Industry_aggregation    32445
Industry_code           32445
Industry_name           32445
Units                   32445
Variable_code           32445
Variable_name           32445
Variable_category       32445
Value                   32445
Industry_code_ANZSIC06  32445
dtype: int64
```

In [50]:
```
print(survey.isnull().sum())
```

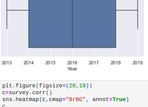
```
Year                    0
Industry_aggregation    0
Industry_code           0
Industry_name           0
Units                   0
Variable_code           0
Variable_name           0
Variable_category       0
Value                   0
Industry_code_ANZSIC06  0
dtype: int64
```

In [51]:
```
survey=survey.dropna()
survey.count()
```

Out[51]:
```
Year                    32445
Industry_aggregation    32445
Industry_code           32445
Industry_name           32445
Units                   32445
Variable_code           32445
Variable_name           32445
Variable_category       32445
Value                   32445
Industry_code_ANZSIC06  32445
dtype: int64
```

In [53]:
```
sns.boxplot(x=survey['Year'])
```

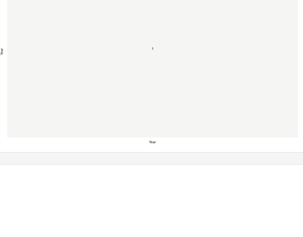
Out[53]:
```
<AxesSubplot:xlabel='Year'>
```



In [66]:
```
plt.figure(figsize=(20,10))
c=survey.corr()
sns.heatmap(c,cmap="BrBG", annot=True)
c
```

Out[66]:

	Year
Year	1.0



In [ ]: