Omair Tariq Capstone Project 1 Milestone Report

Source of Data

Table 4a of the Annual Capital Expenditure Survey for years 2010-2015 from the Census Bureau's website is used for this project [1]. This table details the Capital Expenditures for Structures and Equipment for Companies with Employees by Industry Sector. The table contains how much total money each industry sector spent as well as the amount spent on new and used structures and equipments. North American Industrial Classification System (NAICS) codes [2] are used to identify the Industry Sectors. The tables have roughly 155 Industry Sectors. 55 of these Industry Sectors relate to manufacturing which are the focus of this project. The abbreviations and symbol used in the data are listed below.

Abbreviations and Symbol

The following abbreviations and symbols are used in this publication:

[3]

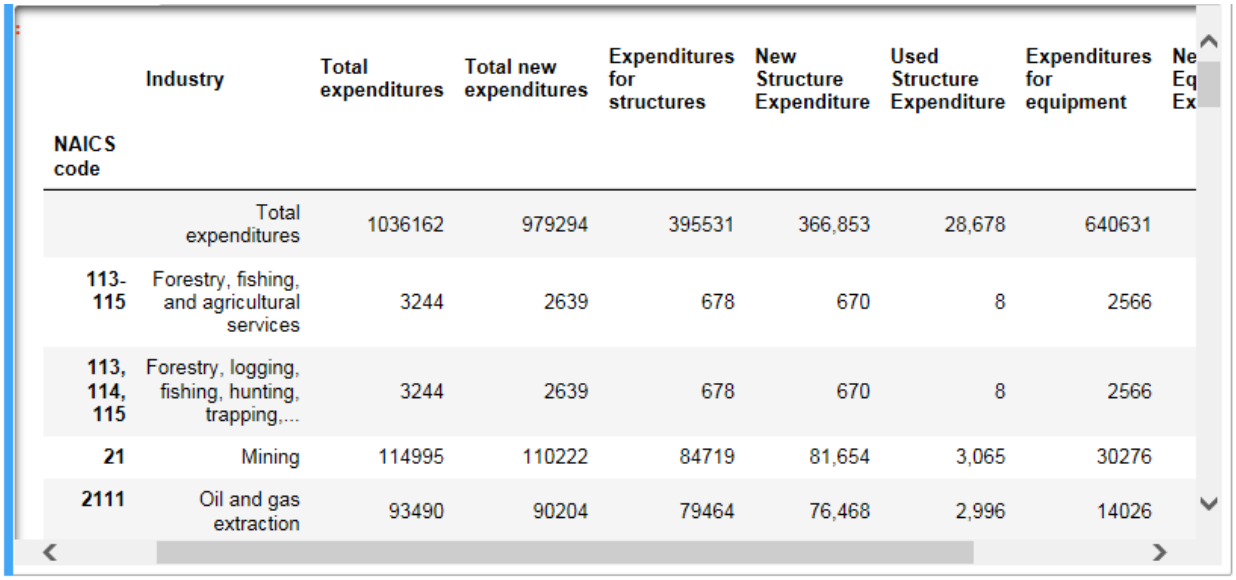
D	The data are withheld (suppressed) to avoid disclosing data for individual enterprises. The data may be included at higher level totals where the individual enterprise's data is not disclosed due to aggregation.
NA	The data are Not Applicable.
NS	The difference in the two estimates being compared are not statistically significant at the Census Bureau standard of 90% confidence. This is also used for the estimate of percent change, where a NS notation means the estimate of percent change is not statistically different from 0%.
X	The estimate of the quality, using Relative Standard Error, is not able to be calculated reliably. This is due to how sampling variance is estimated when most, or all, of the sampled enterprises that had a chance to be sampled did not report positive values for that item.
Z	The estimate is greater than 0, but still rounds to zero in the published units.

Data Wrangling

The tables are in xlsx format and are converted into csv format using excel. Here is a snapshot of the data in csv format :

Table 4a. Capital Expenditures for Structures and Equipment for Companies With Employees by Industry: 2010									
Source : 2010 Annual Capital Expenditures Survey									
See Introduction text for meanings of abbreviations and symbols									
See Appendix A for definition of terms									
See Appendix C for sampling and estimation methodology									
[Millions of current dollars.]									
NAICS cod	Industry	Total expe	Total new	Expenditures for structures			Expenditures for equipment		
				Total	New	Used	Total	New	Used
	Total expenditures	1,036,162	979,294	395,531	366,853	28,678	640,631	612,441	28,190
113-115	Forestry, fishing, and agricultural services	3,244	2,639	678	670	8	2,566	1,969	597
113, 114, 1	Forestry, logging, fishing, hunting, trapping, and agricultural support activities	3,244	2,639	678	670	8	2,566	1,969	597
21	Mining	114,995	110,222	84,719	81,654	3,065	30,276	28,568	1,708
2111	Oil and gas extraction	93,490	90,204	79,464	76,468	2,996	14,026	13,736	290
2121	Coal mining	3,997	3,733	789	780	9	3,208	2,952	256
2122	Metal ore mining	2,499	2,467	1,566	1,558	8	933	909	24
2123	Nonmetallic mineral mining and quarrying	1,982	1,625	585	574	11	1,397	1,051	346
213111, 21	Support activities for oil and gas operations	12,861	12,069	2,304	2,264	40	10,557	9,805	752
213113, 21	Support activities for solid mineral operations	165	124	11	10	1	154	114	40
22	Utilities	94,580	91,645	43,128	42,084	1,044	51,451	49,560	1,891
2211	Electric power generation, transmission, and distribution	83,073	80,538	36,373	35,397	976	46,701	45,142	1,559
2212	Natural gas distribution	8,147	7,807	4,244	(D)	(D)	3,902	(D)	(D)
2213	Water, sewage, and other systems	3,360	3,300	2,511	(D)	(D)	849	(D)	(D)
23	Construction	17,856	13,738	2,633	2,361	272	15,223	11,377	3,846

It can be seen that the first few rows of the datset are not useful for our analysis and are therefore dropped when the csv file is converted into a data frame. It can also be seen that the values have commas in them and this is dealt with on a column-by-column basis. This means that the imported columns are formatted as strings which makes it difficult to perform data analysis. The first attempt to convert the columns is during import by using the ‘thousands’ function in the read_csv command of pandas. Some of the columns are not converted by this method because they contain letter in abbreviations and symbols mentioned above. The second attempt to convert the columns is after looking to see which columns remain to be converted into a float64 datatype.



NAICS code	Industry	Total expenditures	Total new expenditures	Expenditures for structures	New Structure Expenditure	Used Structure Expenditure	Expenditures for equipment	Ne Eq Ex
	Total expenditures	1036162	979294	395531	366,853	28,678	640631	
113-115	Forestry, fishing, and agricultural services	3244	2639	678	670	8	2566	
113, 114, 115	Forestry, logging, fishing, hunting, trapping,...	3244	2639	678	670	8	2566	
21	Mining	114995	110222	84719	81,654	3,065	30276	
2111	Oil and gas extraction	93490	90204	79464	76,468	2,996	14026	

It can be seen above that the columns not converted into type Float64 still have commas in their values. This is because there are some cells in this column with abbreviations and symbols.

The following rows are dropped because they had an abbreviation of (D) in more than one table:

NAICS Code	**Industry**
2212	Natural gas distribution
2213	Water, sewage, and other systems
313,314	Textile mills and textile product mills
315	Apparel manufacturing

The columns which were not imported as type Float 64 are manually converted using the str.replace function in conjunction with the astype(float64) function.There are some cells with Zs or dashed in them which indicates a value of zero. The .replace function is used to convert these values to zero

References

[1] - Annual Capital Expenditure Survey Data Tables
https://www.census.gov/programs-surveys/aces/data/tables.html

[2] – Industry Category Codes List
https://www.census.gov/programs-surveys/aces/information/iccl.html

[3] - Annual Capital Expenditure Survey Abbreviations and Symbols
https://www.census.gov/programs-surveys/aces/about/survey-description.html