Open Canada Data - Analysis of National Defence Contracting and Vendors

In previous analysis of government contract data, the vendor name field was found to be rather unreliable due to a variety of different spellings or misspellings of vendor names. Having to parse through all the names was a bit of a challenge which I did not attempt, however thankfully I found that the Ottawa Civic Tech project had already done much of the hard work on this for me based on an analysis of proactive disclosure data and publicly available for use under an 'unlicence'. (See http://unlicense.org/) Their vendor name information captures many different alternate entries of vendor names and subsidiaries and bundles them under a "parent company". One issue I noted is that parent company level often results in multiple fairly large (from a Canadian perspective), distinct Canadian and foreign-based business units being rolled into one. Depending on the analysis, this kind of grouping of multiple business units, often with very different business lines, may not be ideal.

I have manually updated the vendor data provided by the Ottawa Civic Tech project to include a number of known major and some minor defence suppliers and adjusted parent company name mapping against vendors as a result of recent mergers and acquisitions (for example, Sikorsky is now a business unit of Lockheed Martin). The intent was to improve the quality and tailor it more for an analysis of defence suppliers and the defence industrial base. My updated defence vendor database is publicly available in a csv format on my github repo. In a separate script (wrangling_DND_contracts.R), I imported and wrangled the DND contract data with the cleaned up vendor names joined, into an .rda object. The wrangling script is also available in the repo.

After having integrated vendor name data from the Ottawa Civic Tech project on Government of Canada contract data, I am going to do a short analysis to see if it helps make the analysis of vendor data easier. This analysis was run as of April 2022.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
options(scipen = 999)
load("dnd_contracts.rda")

contract_analysis <- dnd_contracts %>%
  select(vendor_name, contract_date, contract_value,
         economic_object_code, description_en, country_of_vendor,
```

```
         contract_year, parent_company)

summary(contract_analysis)
```

```
##  vendor_name        contract_date        contract_value
##  Length:259312      Min.   :2000-01-01   Min.   :   -471028
##  Class :character   1st Qu.:2010-07-22   1st Qu.:     15692
##  Mode  :character   Median :2014-01-03   Median :     28168
##                     Mean   :2013-12-21   Mean   :    658108
##                     3rd Qu.:2017-12-25   3rd Qu.:     81810
##                     Max.   :2021-12-30   Max.   :4160474985
##  economic_object_code description_en     country_of_vendor  contract_year
##  Length:259312        Length:259312      Length:259312      Length:259312
##  Class :character     Class :character   Class :character   Class :character
##  Mode  :character     Mode  :character   Mode  :character   Mode  :character
##
##
##
##  parent_company
##  Length:259312
##  Class :character
##  Mode  :character
##
##
##
```

There are about 240,000 entries in the data set. The earliest contracts were from the year 2000 while the latest entries reflect contracts let at the end of March 2021. Entry values range from a negative dollar value entry (likely a data entry error or the result of a contract amendment) to $4 billion per single contract award. 3 out of the top 6 of companies receiving contracts from DND were oil companies - clearly fuel contracts are a major source of business!

There are over 180,000 NA's for parent company. That is a lot of entries that are missed even after I addressed some case sensitivity, punctuation, and company suffix issues in the data wrangling.

```
count(contract_analysis, parent_company) %>% arrange(desc(n))
```

```
## # A tibble: 274 x 2
##    parent_company                        n
##    <chr>                             <int>
##  1 <NA>                             191898
##  2 IMPERIAL OIL                       5833
##  3 SIMEX DEFENCE                      3993
##  4 WORLD FUEL SERVICES                3477
##  5 SHELL CANADA PRODUCTS              2980
##  6 JHT                                2915
##  7 CALIAN                             2883
##  8 CANADIAN CORPS OF COMMISSIONAIRES  2029
##  9 UNISOURCE                          1978
## 10 TOP ACES                           1783
## # ... with 264 more rows
```

Most large defence suppliers are identified, however it is still possible many are missed in the 191,000 entries. Below is the list of firms with the largest number of contract awards that are not paired with a parent company.

```
contract_analysis %>%
  filter(is.na(parent_company)) %>%
  count(vendor_name) %>%
  arrange(desc(n))
```

```
## # A tibble: 39,600 x 2
##    vendor_name                      n
##    <chr>                        <int>
##  1 UQSUQ                          735
##  2 TJ NOLAN CONSTRUCTION          611
##  3 ACKLANDS GRAINGER              511
##  4 APRON FUEL SERVICES            473
##  5 RIGHTWAY SANITATION SERVICES   416
##  6 IMPERIAL CLEANERS              401
##  7 AEG FUELS                      379
##  8 CHRYSLER CANADA                372
##  9 LEVITT SAFETY                  330
## 10 MACKINNON AND OLDING           327
## # ... with 39,590 more rows
```

The results are hopeful from this perspective. The firms identified do not appear to be any major contractors. These contracts are undoubtedly for low dollar value and less interesting from defence capability perspective.

Taking another perspective, below are the vendor names doing the largest volume not attributed to a parent company in descending order.

```
contract_analysis %>%
  filter(is.na(parent_company)) %>%
  group_by(vendor_name) %>%
  summarize(contracts_total = sum(contract_value)) %>%
  arrange(desc(contracts_total))
```

```
## # A tibble: 39,600 x 2
##    vendor_name                   contracts_total
##    <chr>                                   <dbl>
##  1 SANTÉ MONTFORT                    521294600
##  2 CANADIAN BORDER OPERATIONS        382800259.
##  3 INDUSTRIES OCÉAN                  200574923.
##  4 EBC                               166725148
##  5 <NA>                              148117169
##  6 NULL                              138375132
##  7 BRONSWERK MARINE                  128496743.
##  8 ZODIAC HURRICANE TECHNOLOGIES     125349760.
##  9 IMPERIAL CLEANERS                 115256635.
## 10 PRIMEX PROJECT MANAGEMENT         104515527.
## # ... with 39,590 more rows
```

The vast majority of contracts in the data base are relatively low value. Relatively speaking, we may not want to spend much time adding in vendors where the overall value is not significant. Lets take a look at the total value of contracts with a parent company identified.

```
parent <- contract_analysis %>% filter(!is.na(parent_company))

a <- sum(parent$contract_value, na.rm = TRUE) #contract value sum with parent
b <- sum(contract_analysis$contract_value, na.rm = TRUE)#contract value all entries

c <- sum(parent$contract_value, na.rm = TRUE)/sum(contract_analysis$contract_value, na.rm = TRUE)

d <- nrow(parent)
e <- nrow(contract_analysis)
f <- d/e

contract_value <- c(round(a, 2), round(b, 2), round(c, 2))
n <- c(as.integer(d), as.integer(e), round(f, 2))
y <- c("with_parent", "all", "percentage")
data.frame(contract_value, n, row.names = y)
```

```
##              contract_value         n
## with_parent 139266276485.42  67414.00
## all         170655363210.89 259312.00
## percentage             0.82      0.26
```

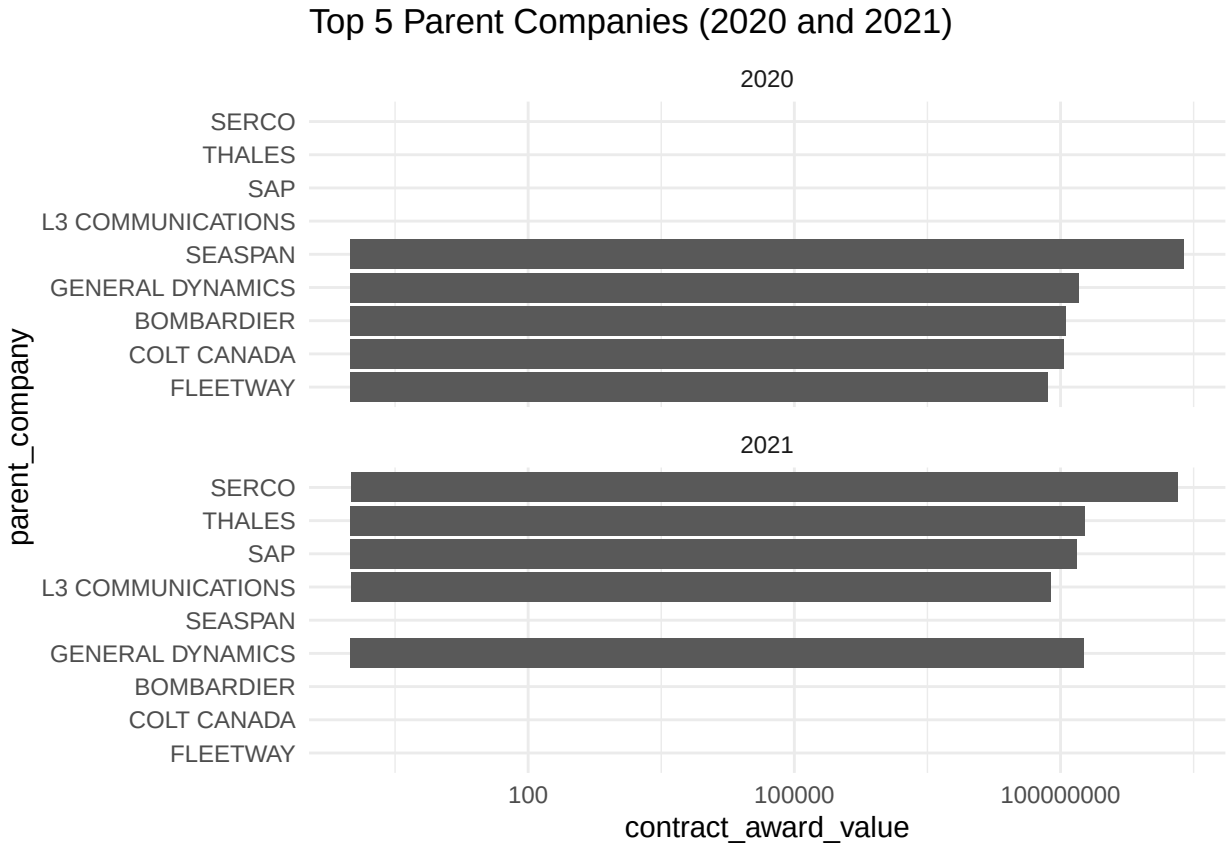
Over 80% of value is captured in the 64,000 some entries. Pretty much all of the largest contracts accounting for several hundred of million or more are attributed to a parent company. I will continue to make updates to the defence vendor name data as time allows, but for now we will live with the 80% solution.

```
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.1.3
```

```
contract_analysis %>% filter(contract_year %in% c("2020","2021"), !is.na(parent_company)) %>%
  group_by(contract_year, parent_company) %>%
  summarize(contract_award_value = sum(contract_value)) %>%
  arrange(desc(contract_award_value)) %>%
  slice_max(order_by = contract_award_value, n=5) %>%
  mutate(parent_company = fct_reorder(parent_company, contract_award_value)) %>%
  ggplot(aes(parent_company, contract_award_value)) + geom_col() +
  scale_y_log10() +coord_flip() +
  facet_wrap(vars(contract_year), nrow = 3) +
  ggtitle("Top 5 Parent Companies (2020 and 2021)") + theme_minimal()
```

```
## `summarise()` has grouped output by 'contract_year'. You can override using the
## `.groups` argument.
```



In 2021, we see a variety of contracts with different providers. SERCO is a services provider, so that contract may have to do with relocation or facility services. General Dynamics represents a number of major defence business units in Canada, however a new contract for armoured vehicles was awarded to General Dynamics Land Systems Canada was awarded in 2020.

1250 and 1251 are the economic object codes for aircraft and aircraft parts respectively. Let's see what who are the biggest suppliers here. Hopefully, there will be no surprises.

```
contract_analysis %>%
  filter(economic_object_code %in% c("1250", "1251"),
         contract_year %in% c("2017", "2018", "2019", "2020", "2021")) %>%
  group_by(contract_year, parent_company, economic_object_code) %>%
  summarize(contract_awards = sum(contract_value)) %>%
  arrange(desc(contract_awards))
```

```
## `summarise()` has grouped output by 'contract_year', 'parent_company'. You can
## override using the `.groups` argument.
```

```
## # A tibble: 45 x 4
## # Groups:   contract_year, parent_company [41]
##    contract_year parent_company                 economic_object~ contract_awards
##    <chr>         <chr>                          <chr>                      <dbl>
##  1 2020          BOMBARDIER                     1251                   116111279.
##  2 2020          UNITED STATES DEPARTMENT OF T~ 1251                    40377732.
##  3 2021          UNITED STATES DEPARTMENT OF T~ 1251                    17064740
##  4 2019          DEW ENGINEERING                1250                    15189635.
```

```
##  5 2018          UNITED STATES DEPARTMENT OF T~ 1251                  13474755
##  6 2017          UNITED STATES DEPARTMENT OF T~ 1251                  13369000
##  7 2019          UNITED STATES DEPARTMENT OF T~ 1251                  13144100
##  8 2020          <NA>                           1251                   7426406.
##  9 2019          <NA>                           1251                   7357406.
## 10 2020          SIMEX DEFENCE                  1251                   7002283.
## # ... with 35 more rows
```

Most of the names are not surprising though I am not familiar with JHT or SIMEX defence, though they figured prominently in the vendor database. However, there are still some very large NA contract award entries under parent company.

Let's try something similar for the Navy with the codes for ships (1256) and ship repair (1257).

```
contract_analysis %>%
  filter(economic_object_code %in% c("1256", "1257"),
         contract_year %in% c("2017", "2018", "2019", "2020", "2021")) %>%
  group_by(contract_year, parent_company, economic_object_code) %>%
  summarize(contract_awards = sum(contract_value)) %>%
  arrange(desc(contract_awards))
```

```
## `summarise()` has grouped output by 'contract_year', 'parent_company'. You can
## override using the `.groups` argument.
```

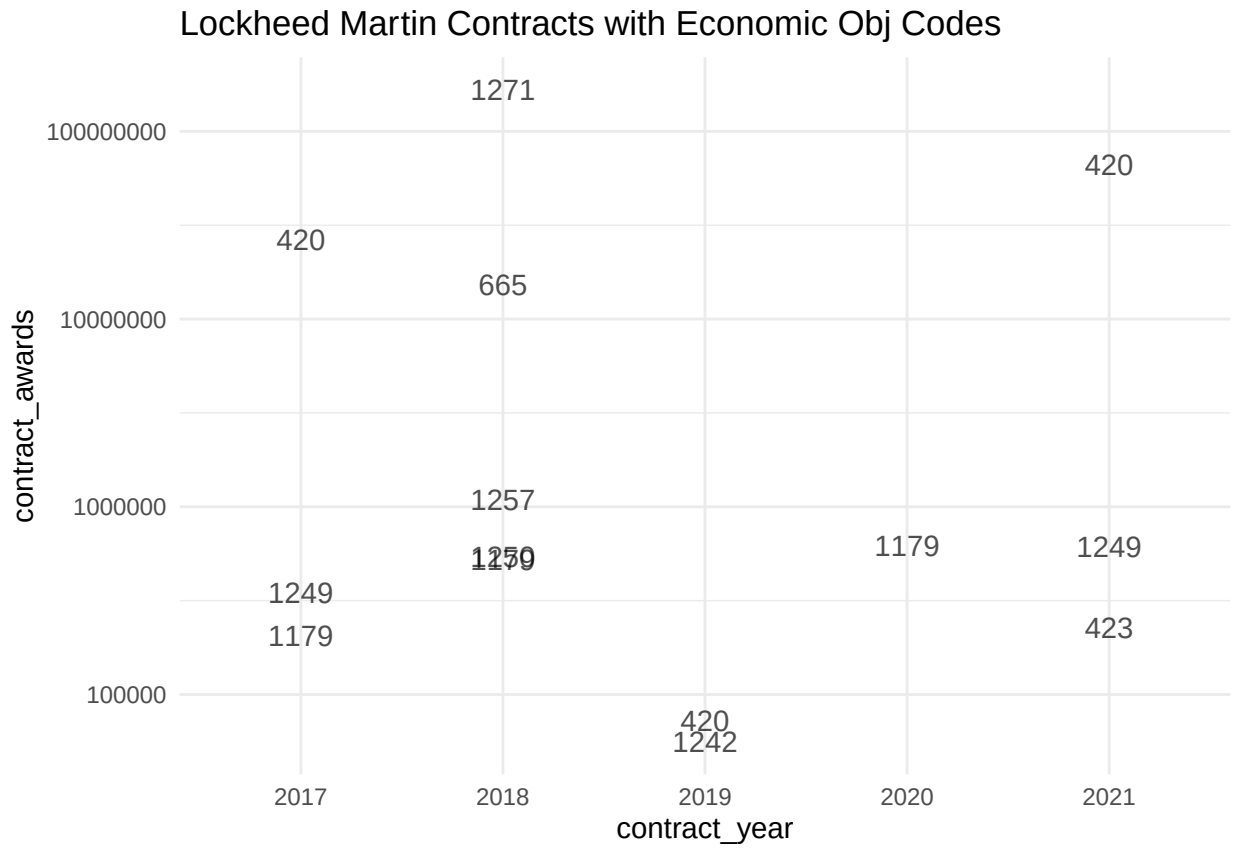
```
## # A tibble: 62 x 4
## # Groups:   contract_year, parent_company [42]
##    contract_year parent_company                 economic_object~ contract_awards
##    <chr>         <chr>                          <chr>                      <dbl>
##  1 2020          SEASPAN                        1257                  2447262476
##  2 2018          SEASPAN                        1256                   322839480.
##  3 2019          <NA>                           1256                   263253753.
##  4 2020          <NA>                           1256                    78002741.
##  5 2017          <NA>                           1257                    29337060.
##  6 2018          <NA>                           1257                    17963515.
##  7 2021          <NA>                           1256                    14157600.
##  8 2018          UNITED STATES DEPARTMENT OF T~ 1257                    13801305
##  9 2017          <NA>                           1256                    12949444.
## 10 2019          <NA>                           1257                    12173730.
## # ... with 52 more rows
```

Again, some very notable NA entries. I also notice that 3M and Bombardier are listed in the table. That seems off giving the ship acquisition and repair coding. It is more likely something was mislabeled.

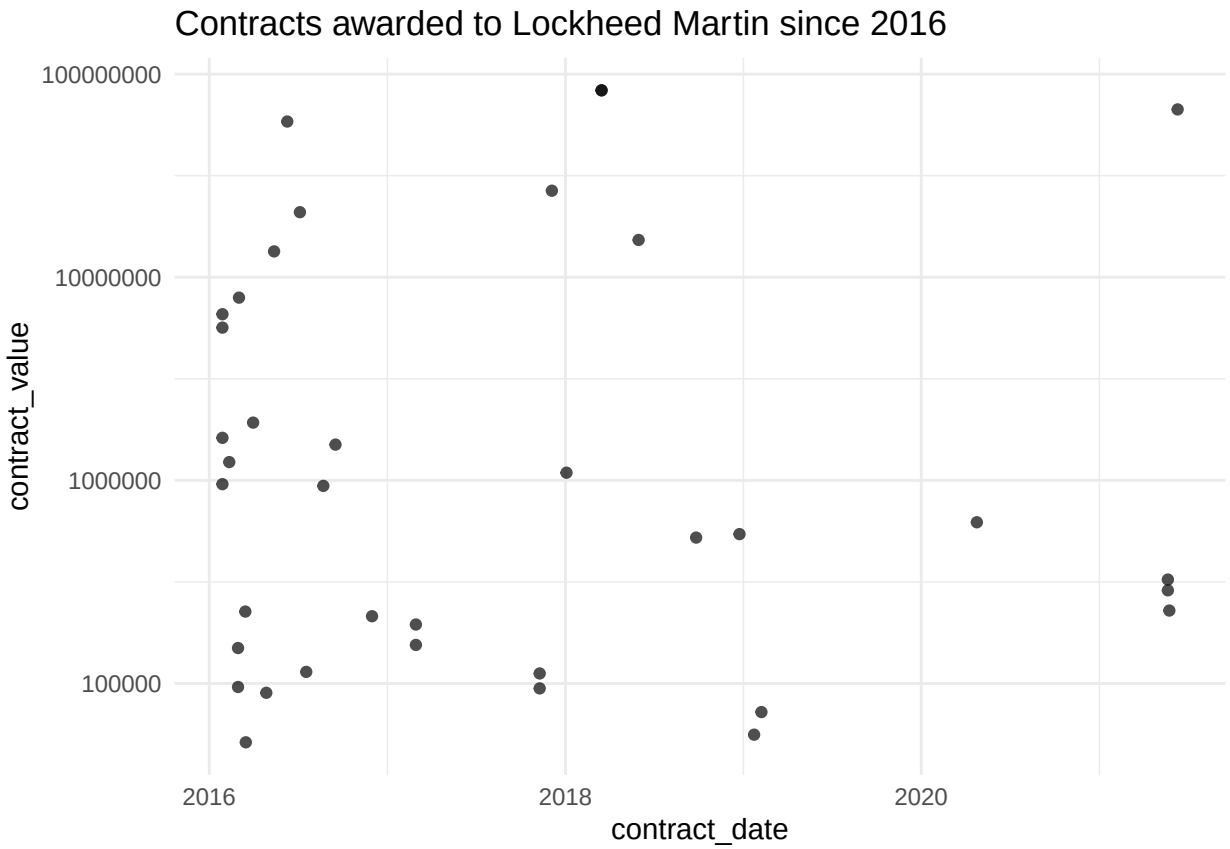
Let's do some specific firm analysis before we wrap this up.

```
contract_analysis %>%
  filter(contract_year %in% c("2017", "2018", "2019", "2020", "2021"),
         parent_company == "LOCKHEED MARTIN") %>%
  group_by(contract_year, parent_company, economic_object_code) %>%
  summarize(contract_awards = sum(contract_value)) %>%
  arrange(desc(contract_awards)) %>%
  ggplot(aes(contract_year, contract_awards, label=economic_object_code)) +
  geom_text(alpha = .7) +scale_y_log10() +
  ggtitle("Lockheed Martin Contracts with Economic Obj Codes") +
  theme_minimal()
```

```
## `summarise()` has grouped output by 'contract_year', 'parent_company'. You can
## override using the `.groups` argument.
```




```
contract_analysis %>%
  filter(parent_company == "LOCKHEED MARTIN", contract_date>"2016-01-01") %>%
  ggplot(aes(contract_date, contract_value)) +
  geom_point(alpha = .7) +
  scale_y_log10() +
  ggtitle("Contracts awarded to Lockheed Martin since 2016") +
  theme_minimal()
```



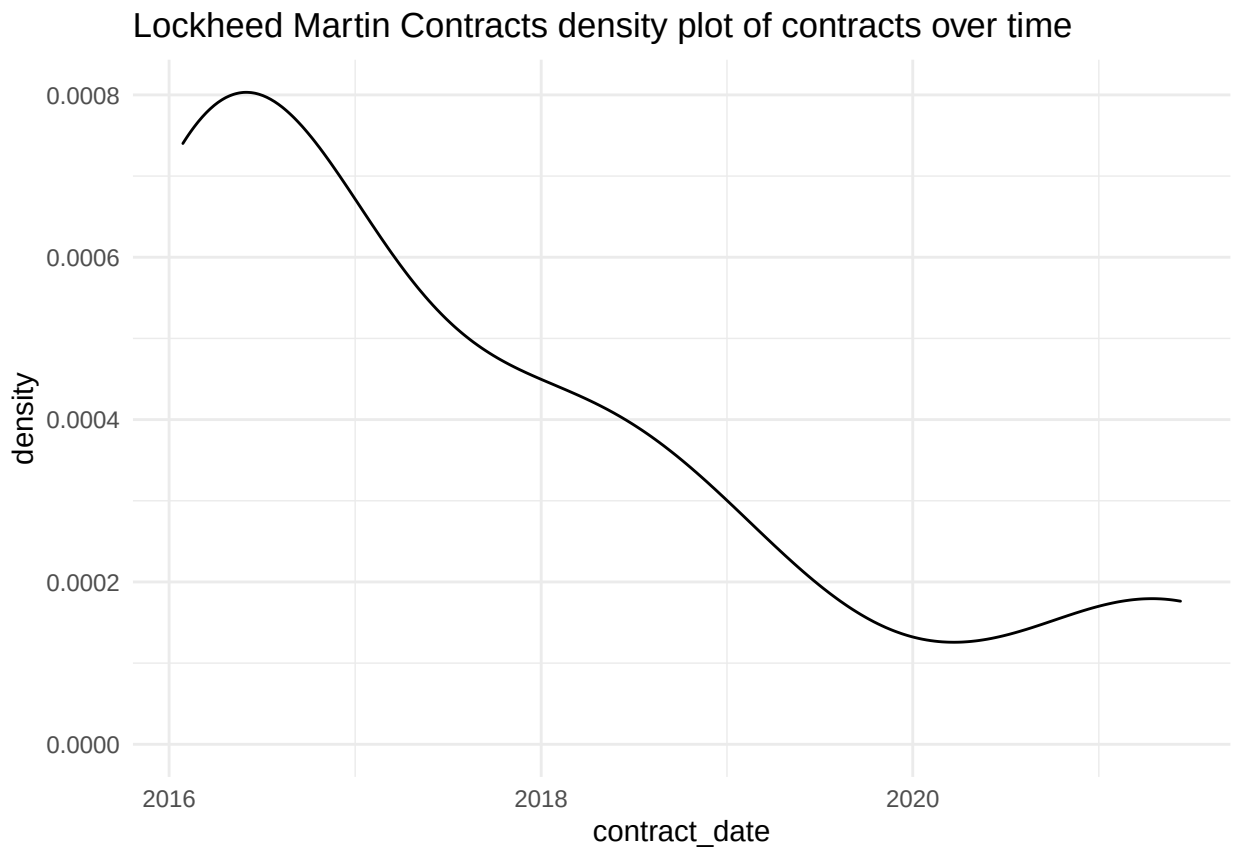
There is a far greater number of points when you do not use the economic object code. I suspect there are a lot of NAs causing for many entries. Another interesting aspect is that this does not capture a significant recent expenditures to Lockheed Martin Canada in the context of shipbuilding. Those contracts flowed through another company so are not reflected here. Clearly the contract data does not capture all the complexities of some of these business relationships.

```
sum(is.na(contract_analysis$economic_object_code))/nrow(contract_analysis)
```

```
## [1] 0.678129
```

About 3/4 of entries are missing their economic object code. Combined with some suspicious entries, I don't think any meaningful analysis using economic object codes in the contract data is possible.

```
contract_analysis %>%
  filter(parent_company == "LOCKHEED MARTIN", contract_date>"2016-01-01") %>%
  ggplot(aes(contract_date)) +
  geom_density() +
  ggtitle("Lockheed Martin Contracts density plot of contracts over time") +
  theme_minimal()
```

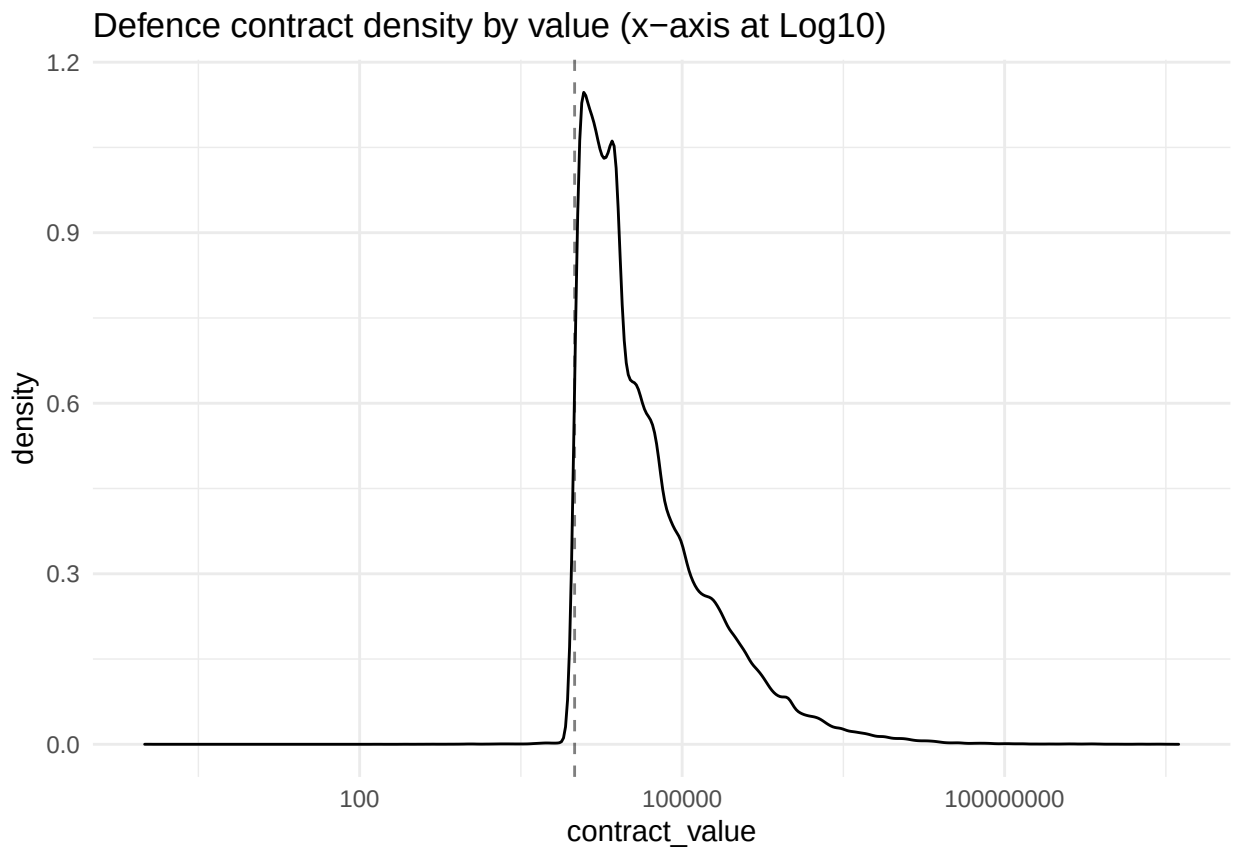



```
contract_analysis %>%
  ggplot(aes(contract_value)) +geom_density() +
  scale_x_log10() +
  ggtitle("Defence contract density by value (x-axis at Log10)") +
  geom_vline(xintercept = 10000, alpha=.5, linetype=2)+
  theme_minimal()
```

```
## Warning in self$trans$transform(x): NaNs produced
```

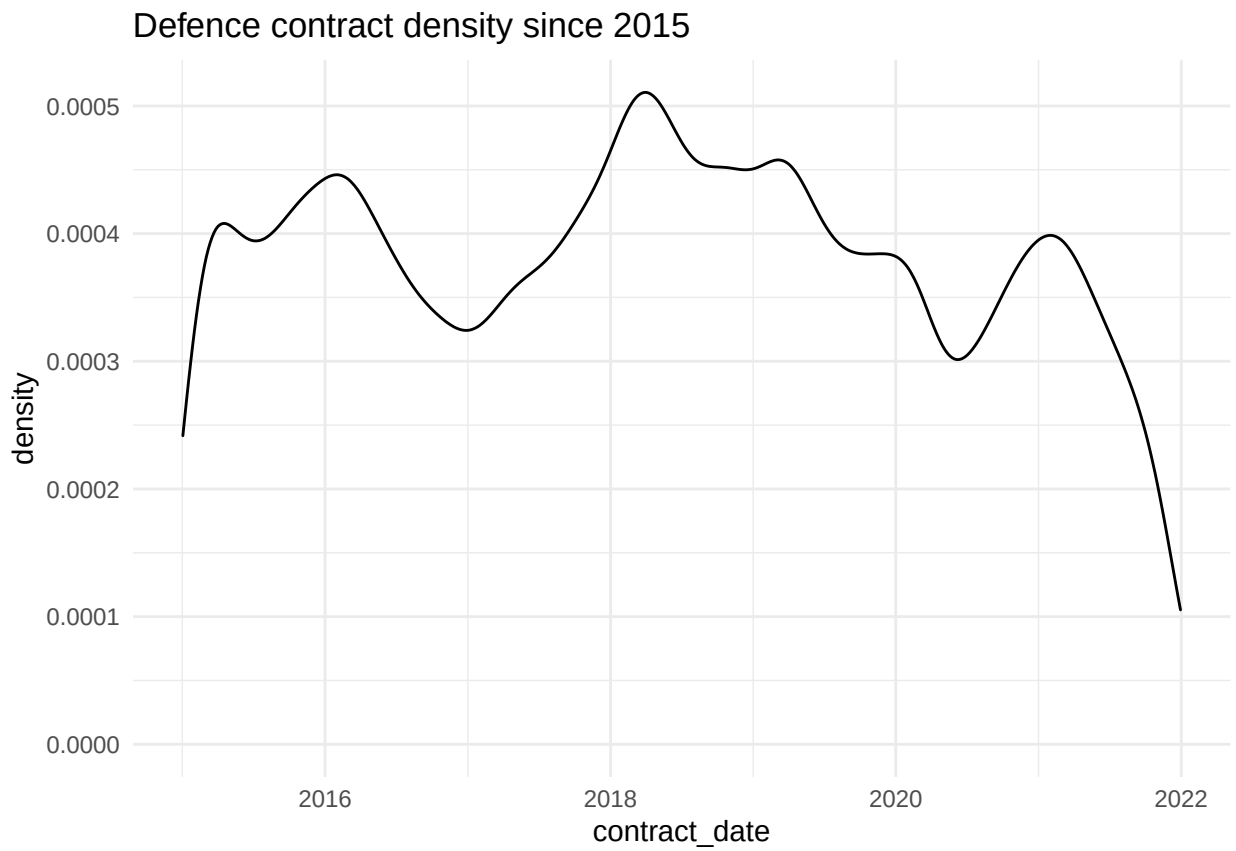
```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 6690 rows containing non-finite values (stat_density).
```



As we can see from the graph, at the 10K mark (noted by the dashed vertical line), the contract entries shoot up. This is logical as this database is only for contracts valued over $10K. We can also see that even with a logarithmic y axis that there is a steep drop in the number of contracts as contract value increases. Using an empirical cumulative distribution function we can see that almost 80% of contracting activity is below $100,000 in value. In fact, almost 99% of defence department contracting activity is below $5 million dollars. This contracting activity would include call ups on standing offers and other contractual arrangements that would be routine and transactional, however it is impressive nonetheless. It also highlights that the most talked about defence contracts in Parliament or in the media only make up a small percentage of the total volume of activity.

```
contract_analysis %>% filter(contract_date>"2015-01-01") %>%
  ggplot(aes(contract_date)) +
  geom_density() +
  ggtitle("Defence contract density since 2015") +
  theme_minimal()
```



As we can see since 2010 there has been a slight drop in the overall volume of contract activity but with with some variation throughout each year. We can likely attribute the peak after 2010 for contracting activity during and towards the end of Canada's mission in Afghanistan. We can see a dip around the 2015 election and the lead up to the release of the 2017 defence policy, however there seems to be growth since that time. I would attribute the drop off of the chart around 2020 to the fact that entries may not be up to date, and the onset of COVID-19 may have caused some data entry delays, even though there are entries in the database as l. We will have to see if that is actually a trend or whether the database just needs to catch up to actual activity.

We will look to update this analysis from time to time.