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 November 2020.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- 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:238136      Min.   :2000-01-01   Min.   :   -471028
##  Class :character   1st Qu.:2010-03-17   1st Qu.:     15776
##  Mode  :character   Median :2013-04-18   Median :     28455
##                     Mean   :2013-05-08   Mean   :    619726
##                     3rd Qu.:2016-12-22   3rd Qu.:     83203
##                     Max.   :2020-09-30   Max.   :4160474985
##
##  economic_object_code description_en     country_of_vendor  contract_year
##  Min.   : 210.0       Length:238136      Length:238136      Length:238136
##  1st Qu.: 630.0       Class :character   Class :character   Class :character
##  Median :1123.0       Mode  :character   Mode  :character   Mode  :character
##  Mean   : 921.4
##  3rd Qu.:1179.0
##  Max.   :6299.0
##  NA's   :175654
##  parent_company
##  Length:238136
##  Class :character
##  Mode  :character
##
##
##
##
```

There are about 225,000 entries in the data set. The earliest contracts were from the year 2000 while the latest entries reflect contracts let at the end of Sept 2020. 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 170,000 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: 273 x 2
##    parent_company                          n
##    <chr>                               <int>
##  1 <NA>                               175512
##  2 IMPERIAL OIL                         5344
##  3 SIMEX DEFENCE                        3852
##  4 WORLD FUEL SERVICES                  3109
##  5 CALIAN                               2819
##  6 JHT                                  2758
##  7 SHELL CANADA PRODUCTS                2641
##  8 UNISOURCE                            1909
##  9 CANADIAN CORPS OF COMMISSIONAIRES    1836
## 10 TOP ACES                             1782
## # ... with 263 more rows
```

Most large defence suppliers are identified, however it is still possible many are missed in the 170,000 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: 37,110 x 2
##    vendor_name                       n
##    <chr>                         <int>
##  1 UQSUQ                           638
##  2 TJ NOLAN CONSTRUCTION           548
##  3 ACKLANDS GRAINGER               482
##  4 APRON FUEL SERVICES             462
##  5 IMPERIAL CLEANERS               391
##  6 RIGHTWAY SANITATION SERVICES    373
##  7 CHRYSLER CANADA                 372
##  8 LEVITT SAFETY                   316
##  9 KAYCOM                          305
## 10 STAFFORD PLUMBING AND HEATING   301
## # ... with 37,100 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))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 37,110 x 2
##    vendor_name                                              contracts_total
##    <chr>                                                              <dbl>
##  1 <NA>                                                       148117169
##  2 NULL                                                       138375132
##  3 ALLIED WINGS LIMITED PARTNERSHIP                           102552318.
##  4 INDUSTRIES OCÉAN                                           101774247
##  5 UTILITIES KINGSTON                                         100294903.
##  6 EODC ENGINEERINGDEVELOPING AND LICENCING                    99308200
##  7 PRIMEX PROJECT MANAGEMENT                                   95372775.
##  8 LLOYDS REGISTER CANADA                                      93643664.
##  9 PYLON ELECTRONICS                                           92724064.
## 10 AVEOS FLEET PERFORMANCE INCAVEOS PERFORMANCE AERONAUTIQUE   90946433
## # ... with 37,100 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.

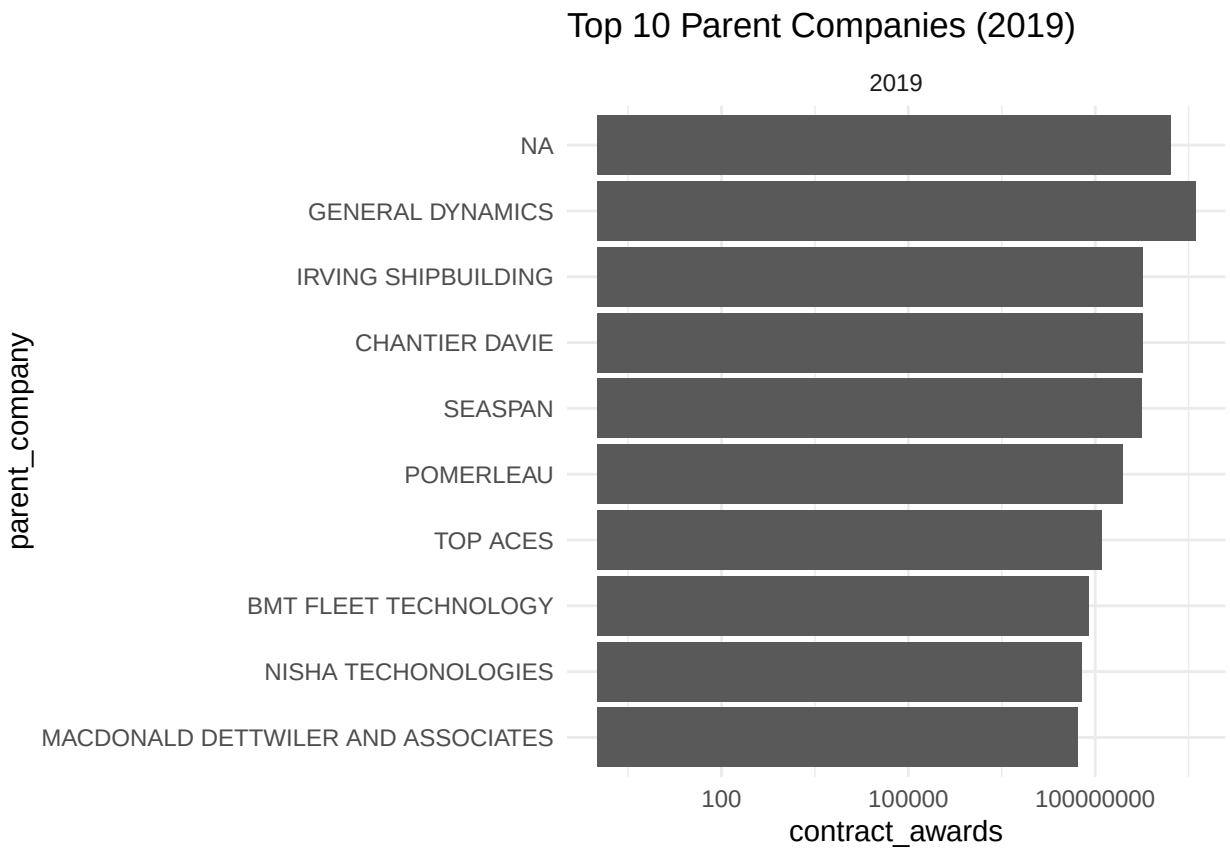
```
parent <- contract_analysis %>% filter(!is.na(parent_company))

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 120420004181.56  62624.00
## all         147579126989.51 238136.00
## percentage              0.82      0.26
```

Over 80% of value is captured in the 60,000 some entries. Pretty much all of the large contracts for billions and hundreds of millions of dollars 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)
contract_analysis %>% filter(contract_year %in% c("2019")) %>%
  group_by(contract_year, parent_company) %>%
  summarize(contract_awards = sum(contract_value)) %>%
  arrange(desc(contract_awards)) %>%
  slice_max(order_by = contract_awards, n=10) %>%
  mutate(parent_company = fct_reorder(parent_company, contract_awards)) %>%
  ggplot(aes(parent_company, contract_awards)) + geom_col() +
  scale_y_log10() +coord_flip() +
  facet_wrap(vars(contract_year), nrow = 3) +
  ggtitle("Top 10 Parent Companies (2019)") + theme_minimal()
```

```
## `summarise()` regrouping output by 'contract_year' (override with `.groups` argument)
```



In playing with the data, there are still some big NA entries in there under the parent company. More clean up of the vendor_name database is going to be needed, however will see what kind of analysis we can do here.

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")) %>%
  group_by(contract_year, parent_company, economic_object_code) %>%
  summarize(contract_awards = sum(contract_value)) %>%
  arrange(desc(contract_awards))
```

```
## `summarise()` regrouping output by 'contract_year', 'parent_company' (override with `.groups` argumer
```

```
## # A tibble: 27 x 4
## # Groups:   contract_year, parent_company [24]
##    contract_year parent_company                economic_object_c~ contract_awards
##    <chr>         <chr>                                      <dbl>           <dbl>
##  1 2019          DEW ENGINEERING                             1250       15189635.
##  2 2018          UNITED STATES DEPARTMENT OF~                1251       13474755
##  3 2017          UNITED STATES DEPARTMENT OF~                1251       13369000
##  4 2019          UNITED STATES DEPARTMENT OF~                1251       13144100
##  5 2019          <NA>                                        1251        7357406.
##  6 2017          <NA>                                        1251        4342862.
```

```
##  7 2017          UNITED STATES DEPARTMENT OF~              1251        3161796
##  8 2018          <NA>                                      1251        2615640.
##  9 2017          SIMEX DEFENCE                             1251        2040744.
## 10 2017          JHT                                       1251        1802153.
## # ... with 17 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")) %>%
  group_by(contract_year, parent_company, economic_object_code) %>%
  summarize(contract_awards = sum(contract_value)) %>%
  arrange(desc(contract_awards))
```

```
## `summarise()` regrouping output by 'contract_year', 'parent_company' (override with `.groups` argumer
```

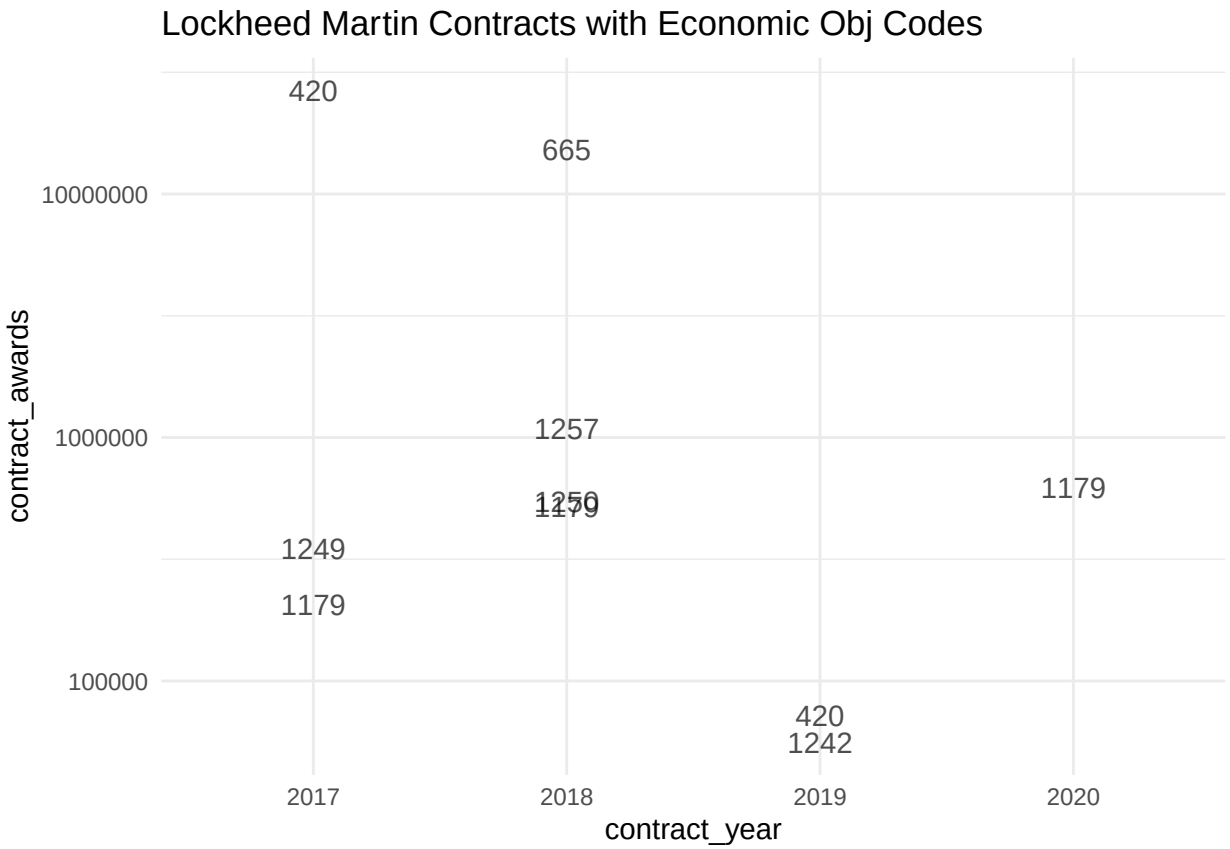
```
## # A tibble: 40 x 4
## # Groups:   contract_year, parent_company [28]
##    contract_year parent_company               economic_object_c~ contract_awards
##    <chr>         <chr>                                   <dbl>           <dbl>
##  1 2018          SEASPAN                                  1256      322839480.
##  2 2019          <NA>                                     1256      148216543.
##  3 2017          <NA>                                     1257       29337060.
##  4 2018          <NA>                                     1257       17963515.
##  5 2018          UNITED STATES DEPARTMENT OF~             1257       13801305
##  6 2017          <NA>                                     1256       12949444.
##  7 2019          <NA>                                     1257       11808039.
##  8 2018          <NA>                                     1256       11512412.
##  9 2018          SIMEX DEFENCE                            1257        6159492.
## 10 2018          THALES                                   1257        2782672.
## # ... with 30 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.

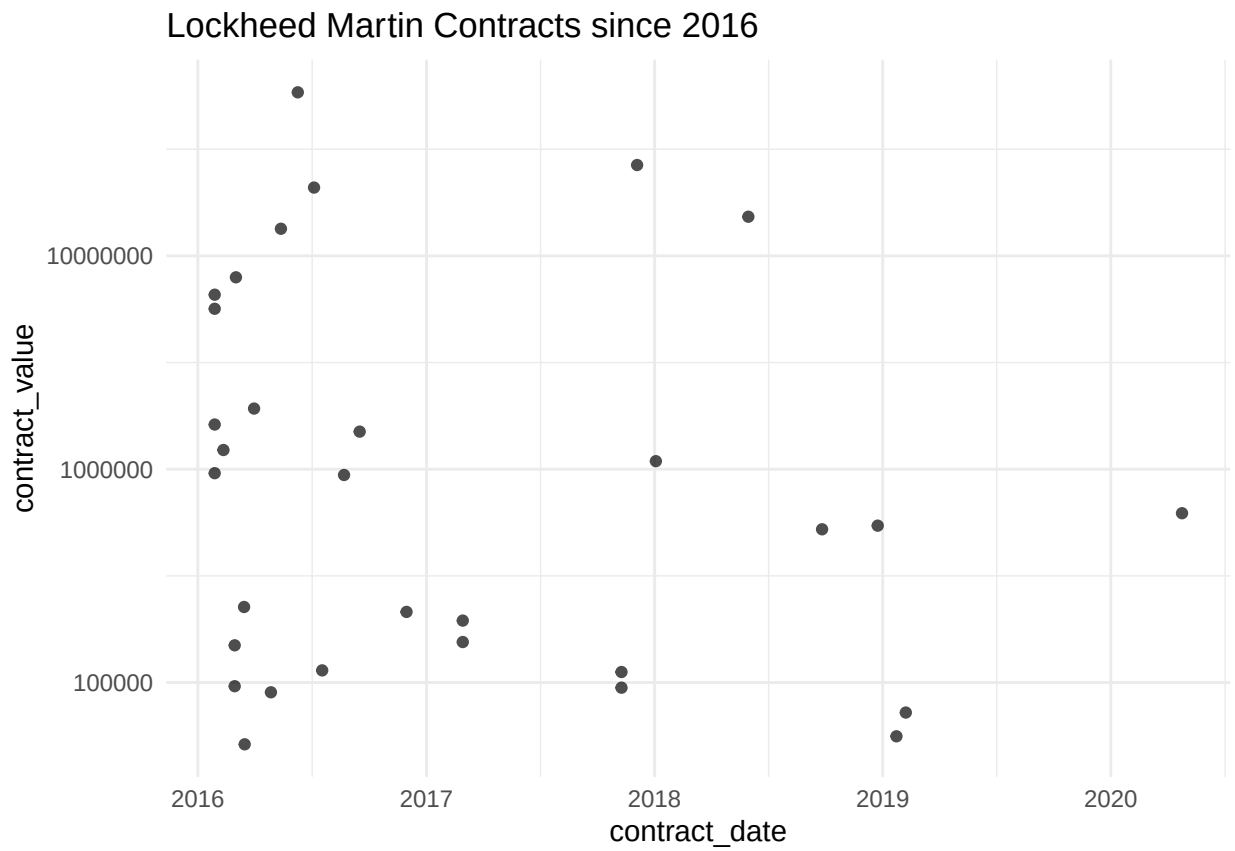
Let's do some specific firm analysis before we wrap this up.

```
contract_analysis %>%
  filter(contract_year %in% c("2017", "2018", "2019", "2020"),
         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()` regrouping output by 'contract_year', 'parent_company' (override with `.groups` argumer
```




```
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("Lockheed Martin Contracts since 2016") +
  theme_minimal()
```



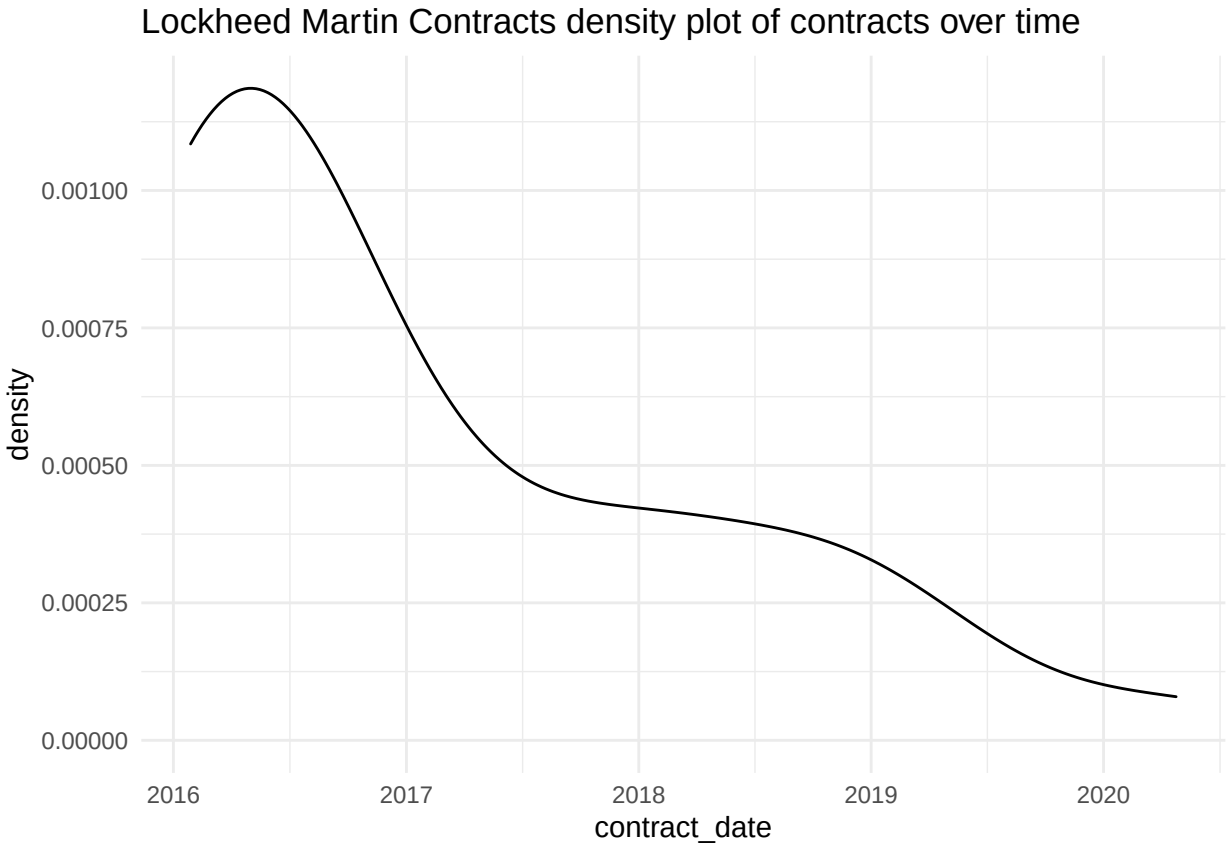
There is a far greater number of points when you do not use the economic object code. I suspect there are a lot of NAs causing for many entries.

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

```
## [1] 0.7376205
```

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()
```

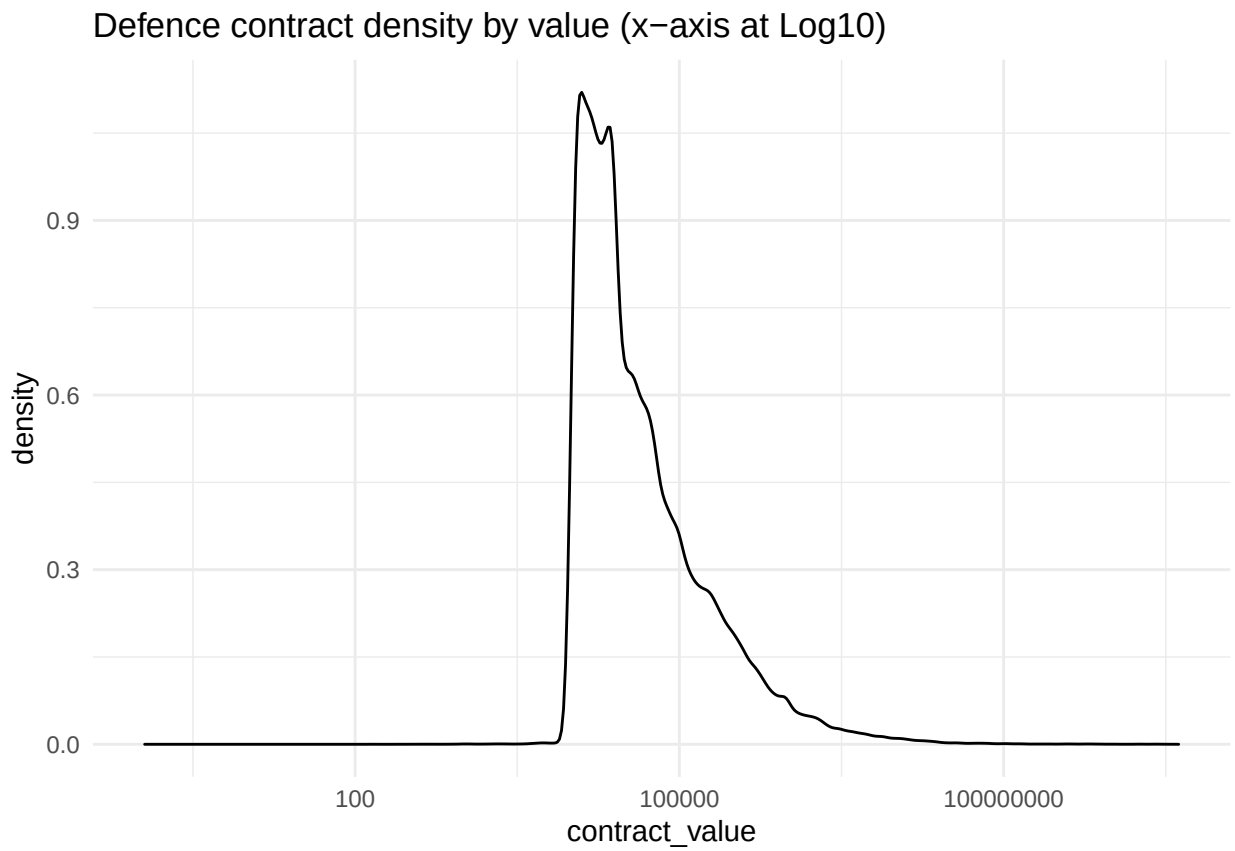



```
contract_analysis %>%
  ggplot(aes(contract_value)) +geom_density() +
  scale_x_log10() +
  ggtitle("Defence contract density by value (x-axis at Log10)") +
  theme_minimal()
```

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

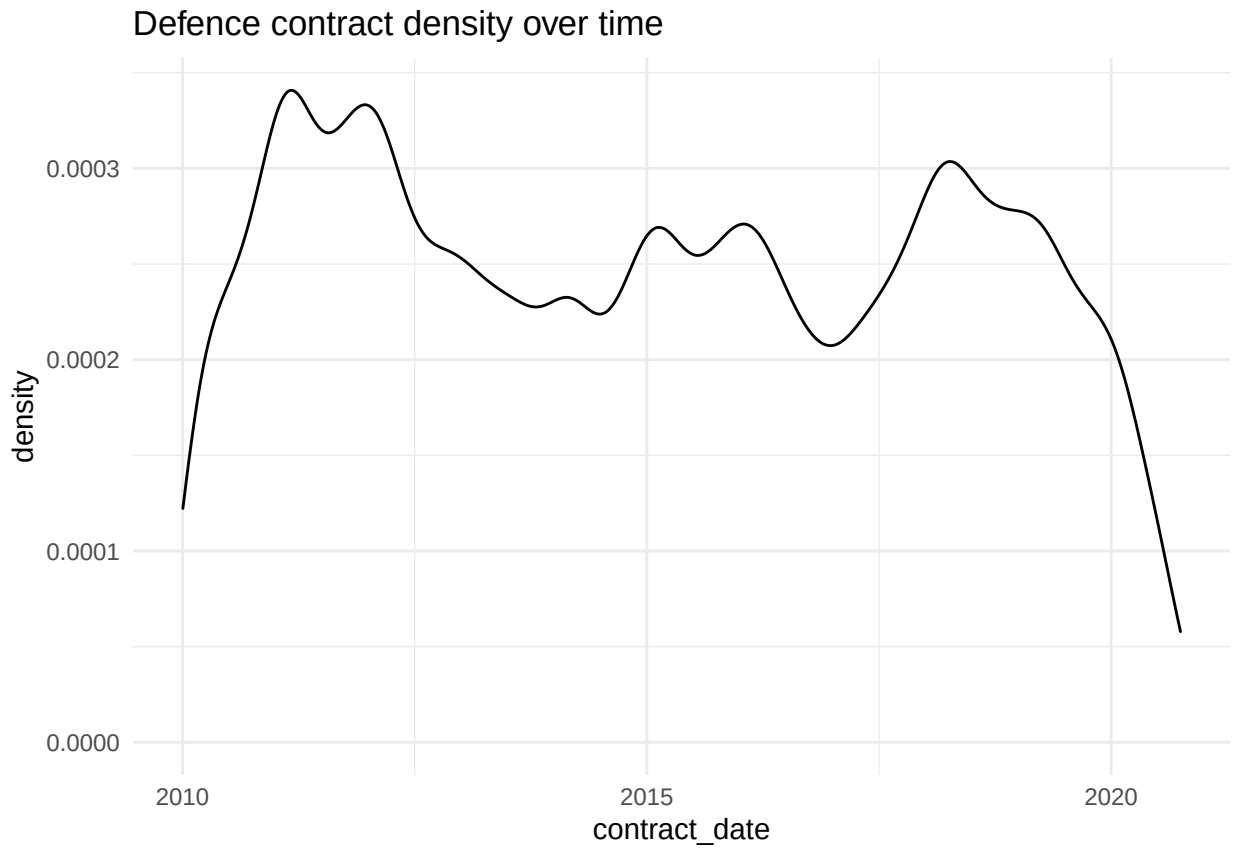
```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 6671 rows containing non-finite values (stat_density).
```



As we can see from the graph, at the 10K mark, 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>"2010-01-01") %>%
  ggplot(aes(contract_date)) +geom_density() +
  ggtitle("Defence contract density over time") +
  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 late as Sept 2020. 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.