Blackduck workshop challenge

the most common license ids

```
data <- read_feather("../results/data.feather")

top_lic <- data %>%
  group_by(license_id) %>%
  summarise(num = n()) %>%
  arrange(desc(num))


head(top_lic,10)
```

```
## # A tibble: 10 × 2
##    license_id    num
##         <int>  <int>
## 1          45 439800
## 2          -1 402164
## 3          34 388921
## 4          30 258647
## 5          11 211177
## 6           3 194419
## 7          36 142244
## 8          13  85854
## 9          12  66346
## 10         35  49832
```

Insight: the most popular projects?

ranked by how many copies are in the dataset for each open source project id

```
(top_uuid <- data %>%
  group_by(d_r_uuid) %>%
  summarise(num = n()) %>%
  arrange(desc(num)) %>%
  head(10))
```

```
## # A tibble: 10 × 2
##                                d_r_uuid   num
##                                  <fctr> <int>
## 1  2ff38a7c-238b-487f-af9b-64a2ec81d81c 46680
## 2  3f28ddca-0c78-4a3a-9e9f-ead210db5368 32571
## 3  e6ec4531-9b98-4653-8f4b-5cb27326d101 22784
## 4  ca691cd3-5ded-46df-935d-2691bbf3c28d 19072
## 5  1a7d76e1-ceca-4da3-911c-c22e4368c32a 13794
## 6  590ef749-4814-4c87-80e5-0f31390a5cdf 13794
## 7  020fb5c7-33e1-4dfd-9e8b-5bb2a3510113 13775
```

```
## 8  5e21fa93-6af9-44ad-b422-f34614db6d1f 13680
## 9  6925dc3d-499b-402b-9a19-6b96de0976d4 13642
## 10 6537a974-ba40-4921-8410-692f1bcf77bc 13205
```

What's the most popular version for each of the top projects?

```
(data %>%
  filter(d_r_uuid %in% top_uuid$d_r_uuid) %>%
  group_by(d_r_uuid,version) %>%
  summarise(num = n()) %>%
  arrange(desc(num)))
```

```
## Source: local data frame [10 x 3]
## Groups: d_r_uuid [10]
##
##                                 d_r_uuid     version   num
##                                   <fctr>      <fctr> <int>
## 1  2ff38a7c-238b-487f-af9b-64a2ec81d81c    8u72-b15 46680
## 2  3f28ddca-0c78-4a3a-9e9f-ead210db5368  7u85-2.6.1 32571
## 3  e6ec4531-9b98-4653-8f4b-5cb27326d101 6b18-1.8.13 22784
## 4  ca691cd3-5ded-46df-935d-2691bbf3c28d      38.5.2 19072
## 5  1a7d76e1-ceca-4da3-911c-c22e4368c32a       5.0.1 13794
## 6  590ef749-4814-4c87-80e5-0f31390a5cdf   5.0.1~rc2 13794
## 7  020fb5c7-33e1-4dfd-9e8b-5bb2a3510113   5.0.1~rc1 13775
## 8  5e21fa93-6af9-44ad-b422-f34614db6d1f       5.0.2 13680
## 9  6925dc3d-499b-402b-9a19-6b96de0976d4       4.4.4 13642
## 10 6537a974-ba40-4921-8410-692f1bcf77bc       4.4.4 13205
```

Note that the number of copies for the "most popular version" of each top project is the same with the total number of that project, as shown in the previous table.

Insight: Each d_r_uuid (open source project id) is associated with a unique version of a project. In other words, different versions of the same project will have different open source project ids.

combine levels based on their occurring frequency in the dataset.

```
data$license_id <- as.factor(data$license_id)

dws_freq <- as.data.frame(table(data$dws))
names(dws_freq) <- c("dws","dwsfreq")

dns_freq <- as.data.frame(table(data$dns))
names(dns_freq) <- c("dns","dnsfreq")

so_freq <- as.data.frame(table(data$so))
names(so_freq) <- c("so","sofreq")

lic_freq <- as.data.frame(table(data$license_id))
names(lic_freq) <- c("license_id","licfreq")
```

```
combined <- inner_join(dws_freq,data,by="dws")
combined <- inner_join(dns_freq,combined,by = "dns")
combined <- inner_join(so_freq,combined,by = "so")
combined <- inner_join(lic_freq,combined,by = "license_id")

data_freq <- combined %>%
  select(d_r_uuid, version, dwsfreq, dnsfreq, sofreq, licfreq)
```

show most common dws

```
combined %>%
  arrange(desc(dwsfreq)) %>%
  select(dws,dns,so,license_id) %>%
  head(10)
```

```
##                                 dws                                 dns
## 1  3724c22825aad4efd8c3e47c6b2b0ee7 d8bd94ceb70d78d28aec9507e71478c1
## 2  3724c22825aad4efd8c3e47c6b2b0ee7 d8bd94ceb70d78d28aec9507e71478c1
## 3  3724c22825aad4efd8c3e47c6b2b0ee7 d8bd94ceb70d78d28aec9507e71478c1
## 4  3724c22825aad4efd8c3e47c6b2b0ee7 d8bd94ceb70d78d28aec9507e71478c1
## 5  3724c22825aad4efd8c3e47c6b2b0ee7 d8bd94ceb70d78d28aec9507e71478c1
## 6  3724c22825aad4efd8c3e47c6b2b0ee7 d8bd94ceb70d78d28aec9507e71478c1
## 7  3724c22825aad4efd8c3e47c6b2b0ee7 d8bd94ceb70d78d28aec9507e71478c1
## 8  3724c22825aad4efd8c3e47c6b2b0ee7 d8bd94ceb70d78d28aec9507e71478c1
## 9  3724c22825aad4efd8c3e47c6b2b0ee7 d8bd94ceb70d78d28aec9507e71478c1
## 10 3724c22825aad4efd8c3e47c6b2b0ee7 d8bd94ceb70d78d28aec9507e71478c1
##                                  so license_id
## 1  d41d8cd98f00b204e9800998ecf8427e         -1
## 2  d41d8cd98f00b204e9800998ecf8427e         -1
## 3  d41d8cd98f00b204e9800998ecf8427e         -1
## 4  d41d8cd98f00b204e9800998ecf8427e         -1
## 5  d41d8cd98f00b204e9800998ecf8427e         -1
## 6  d41d8cd98f00b204e9800998ecf8427e         -1
## 7  d41d8cd98f00b204e9800998ecf8427e         -1
## 8  d41d8cd98f00b204e9800998ecf8427e         -1
## 9  d41d8cd98f00b204e9800998ecf8427e         -1
## 10 d41d8cd98f00b204e9800998ecf8427e         -1
```

show most common dns

```
combined %>%
  arrange(desc(dnsfreq)) %>%
  select(dws,dns,so,license_id) %>%
  head(10)
```

```
##                                 dws                                 dns
## 1  00570b154fdad9c586553da348127a11 eef1ddca51de16bb21dcc01fef368f20
## 2  00570b154fdad9c586553da348127a11 eef1ddca51de16bb21dcc01fef368f20
## 3  00570b154fdad9c586553da348127a11 eef1ddca51de16bb21dcc01fef368f20
## 4  00e94551f81cd370301e206b59d283c7 eef1ddca51de16bb21dcc01fef368f20
## 5  00e94551f81cd370301e206b59d283c7 eef1ddca51de16bb21dcc01fef368f20
## 6  00e94551f81cd370301e206b59d283c7 eef1ddca51de16bb21dcc01fef368f20
```

```
## 7  00fa17d35f38b7b0f62203f7f6e5aa3c eef1ddca51de16bb21dcc01fef368f20
## 8  00fa17d35f38b7b0f62203f7f6e5aa3c eef1ddca51de16bb21dcc01fef368f20
## 9  00fa17d35f38b7b0f62203f7f6e5aa3c eef1ddca51de16bb21dcc01fef368f20
## 10 0155415a6ff2cf2e369f2c2e4dbf4a92 eef1ddca51de16bb21dcc01fef368f20
##                                  so license_id
## 1  d41d8cd98f00b204e9800998ecf8427e         11
## 2  d41d8cd98f00b204e9800998ecf8427e         11
## 3  d41d8cd98f00b204e9800998ecf8427e         11
## 4  d41d8cd98f00b204e9800998ecf8427e         11
## 5  d41d8cd98f00b204e9800998ecf8427e         11
## 6  d41d8cd98f00b204e9800998ecf8427e         11
## 7  d41d8cd98f00b204e9800998ecf8427e         11
## 8  d41d8cd98f00b204e9800998ecf8427e         11
## 9  d41d8cd98f00b204e9800998ecf8427e         11
## 10 d41d8cd98f00b204e9800998ecf8427e         11
```

show most common so

```
combined %>%
  arrange(desc(sofreq)) %>%
  select(dws,dns,so,license_id) %>%
  head(10)
```

```
##                                  dws                                dns
## 1  416c0cdeb2667f55765fa0949d8f7f64 0406aa6d2763123fe901118a9cbebafa
## 2  71ea3bf3c2385525d47a352246418f12 041ab0278c47f77fcaa0238d068565b4
## 3  3970cf060e594cda0dcd1afe0bf00740 0cb79a3aa381d0699b3e23e05a01a843
## 4  6ac255e3e0ed1b41ef0ea961129b6b81 0e7817ca02fccb5ea14aefd16d668413
## 5  5f3aeafb82e0de0eff7ecaa90f465175 182418d9944c9c175926decebf75ee3e
## 6  92160ac74d1cc4f8fe585c73fbaa7981 1cc55a85505d765168d0e807613f7edb
## 7  2436cef1fe6d5efb79fa26e7d5543ea7 1e5cf817b6640b147d54c2e2c7b651ab
## 8  29123ec057e2bf4ac0411c8a3e667ecf 1e5cf817b6640b147d54c2e2c7b651ab
## 9  2d3f1aaf15469d8548e70348c35e0ed6 1e5cf817b6640b147d54c2e2c7b651ab
## 10 52b548e1ec7dbe2761e2edd4b6f19ca9 1e5cf817b6640b147d54c2e2c7b651ab
##                                  so license_id
## 1  d41d8cd98f00b204e9800998ecf8427e         -4
## 2  d41d8cd98f00b204e9800998ecf8427e         -4
## 3  d41d8cd98f00b204e9800998ecf8427e         -4
## 4  d41d8cd98f00b204e9800998ecf8427e         -4
## 5  d41d8cd98f00b204e9800998ecf8427e         -4
## 6  d41d8cd98f00b204e9800998ecf8427e         -4
## 7  d41d8cd98f00b204e9800998ecf8427e         -4
## 8  d41d8cd98f00b204e9800998ecf8427e         -4
## 9  d41d8cd98f00b204e9800998ecf8427e         -4
## 10 d41d8cd98f00b204e9800998ecf8427e         -4
```

Insight: All of the most common dws, and dns have a common so.

To examine this further:

```
combined %>%
  group_by(so,dns,dws) %>%
```

```
  summarise(num = n()) %>%
  arrange(desc(num)) %>%
  nrow()
```

```
## [1] 1898658
```

```
nlevels(as.factor(data$dws))
```

```
## [1] 1898658
```

```
combined %>%
  group_by(so,dns) %>%
  summarise(num = n()) %>%
  arrange(desc(num)) %>%
  nrow()
```

```
## [1] 1438609
```

```
nlevels(as.factor(data$dns))
```

```
## [1] 1438609
```

```
combined %>%
  group_by(so) %>%
  summarise(num = n()) %>%
  arrange(desc(num)) %>%
  nrow()
```

```
## [1] 578502
```

```
nlevels(as.factor(data$so))
```

```
## [1] 578502
```

Insight: There is a hierachical relationship among the features. Two projects with the same dws cannot have different dns or so's. Two projects with the same dns cannot have different so's.

The interpretation of this could be that so is the socket number on a server level. One so can be connected to many dns, which is the domain name services on an ip level. Finally, each ip can be associated with multiple dws.

Knowing this, the same project tend to have the same dws. Different projects can be connected by the same dns, or the same so.

Insight: the "largest"" project with the most common dws is:

```
top_dws <- combined %>%
  arrange(desc(dwsfreq)) %>%
  select(dws,dwsfreq)

head(top_dws,1)
```

```
##                                 dws dwsfreq
## 1 3724c22825aad4efd8c3e47c6b2b0ee7    1152
```

it has 33 versions:

```
combined %>%
  filter(dws == top_dws$dws[1]) %>%
  group_by(version) %>%
  summarise(num = n()) %>%
  arrange(desc(num))
```

```
## # A tibble: 33 × 2
##      version   num
##       <fctr> <int>
## 1      5.0.1   152
## 2      4.4.4   114
## 3      4.2.5    95
## 4      4.3.3    79
## 5  4.4.4~rc1    76
## 6      5.0.2    76
## 7      4.4.2    57
## 8      4.4.5    57
## 9  5.0.1~rc1    57
## 10     4.2.6    38
## # ... with 23 more rows
```

and used 23 licenses

```
combined %>%
  filter(dws == top_dws$dws[1]) %>%
  group_by(license_id) %>%
  summarise(num = n()) %>%
  arrange(desc(num))
```

```
## # A tibble: 23 × 2
##    license_id   num
##        <fctr> <int>
## 1          11    59
## 2          12    59
## 3          13    59
## 4          22    59
## 5          30    59
## 6          34    59
## 7          53    59
## 8          87    59
```

```
## 9          88    59
## 10        121    59
## # ... with 13 more rows
```