Compare missing versus non-missing Statistics*

Deal and Lawyer Studies

Li Yuan, li.yuan@vanderbilt.edu

26 July, 2021 17:12:46

Abstract

The aim of this paper is to study correlation between missing information and industry sector, year, age, deal value. We will focus on how gender affects missing info using machine learning inference in the second half of this paper.

Contents

*Advisor: Tracey George and Albert

Load packages

```
> library(readtext)
> library(antiword)
> library(tidyverse)
> library(ggplot2)
> library(textreadr)
> library(stringi)
> library(textclean)
> library(SemNetCleaner)
> library(readxl)
>
> library(patchwork)
> library(ggrepel)
> library(gghighlight)
> library(paletteer)
> library(ggExtra)
> library(ggbeeswarm)
> library(kableExtra)
> library(caret)
> library(randomForest)
> library(corrplot)
> library(lubridate)
```

Load data

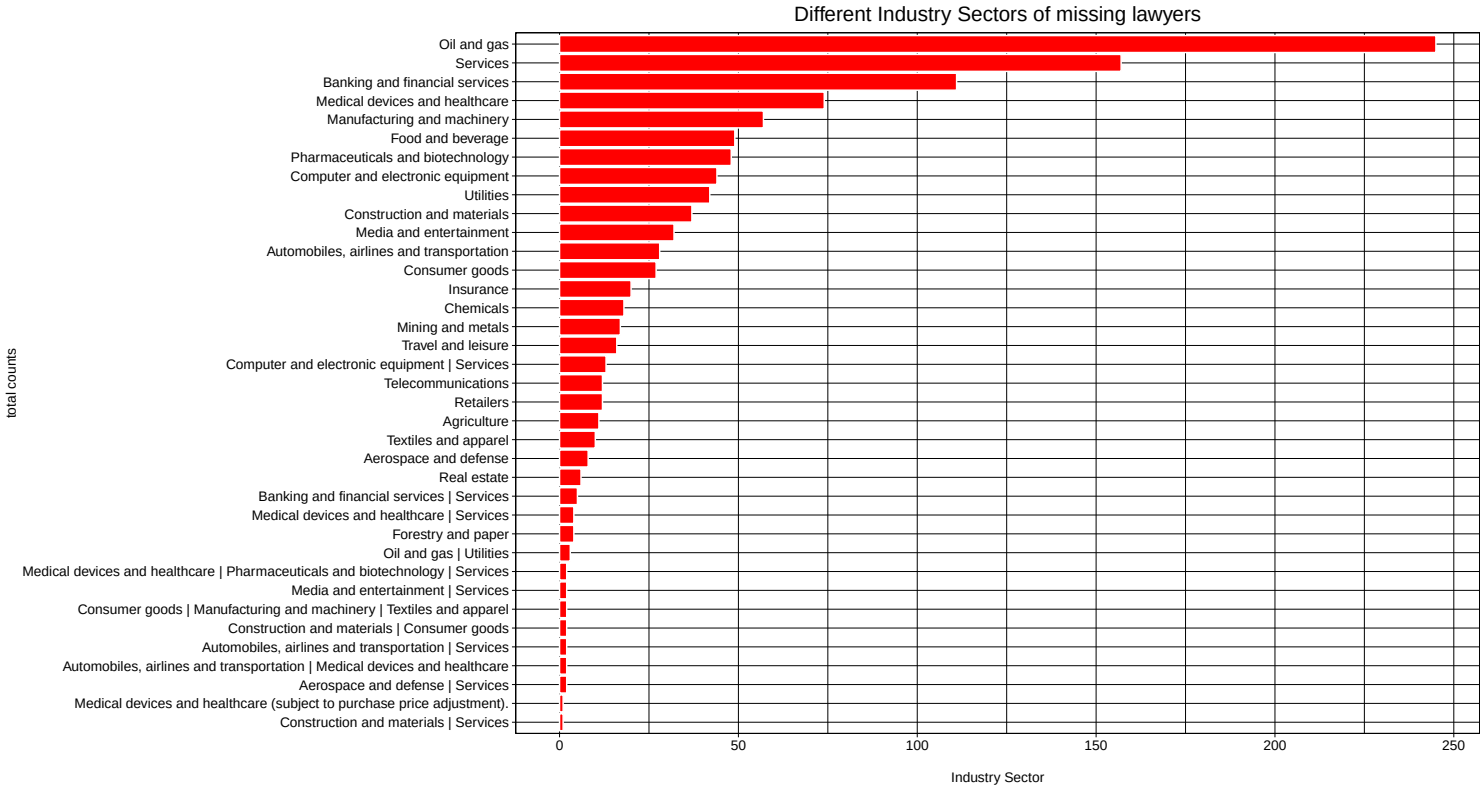
```
> deal <- read_csv("../data/deal/deal(1).csv", col_types = cols(.default = "c"))
> # View(deal)
>
> merge_deal_lawyer <- read_csv("../data/deal_lawyer/merge_deal_lawyer.csv")
> # View(merge_deal_lawyer)
>
> distin_com_lawyer <- read_csv("../data/lawyer/keep_MA_first.csv", col_types = cols(.default = "c"))
> # View(distin_com_lawyer)
>
> map_index <- read_csv("../data/deal/map_index.csv")
> # View(map_index)
>
> encoded_merge_dl <- merge_deal_lawyer %>%
+     left_join(map_index, by = c(deal = "Deal_name")) %>%
+     select(Deal_number, everything(), -deal)
> # View(encoded_merge_dl)
>
> encoded_deal <- deal %>%
+     left_join(map_index, by = c(`Deal name` = "Deal_name")) %>%
+     select(Deal_number, everything(), -`Deal name`)
> # View(encoded_deal)
```

How many lawyers are unknown (NA and "not disclosed")?

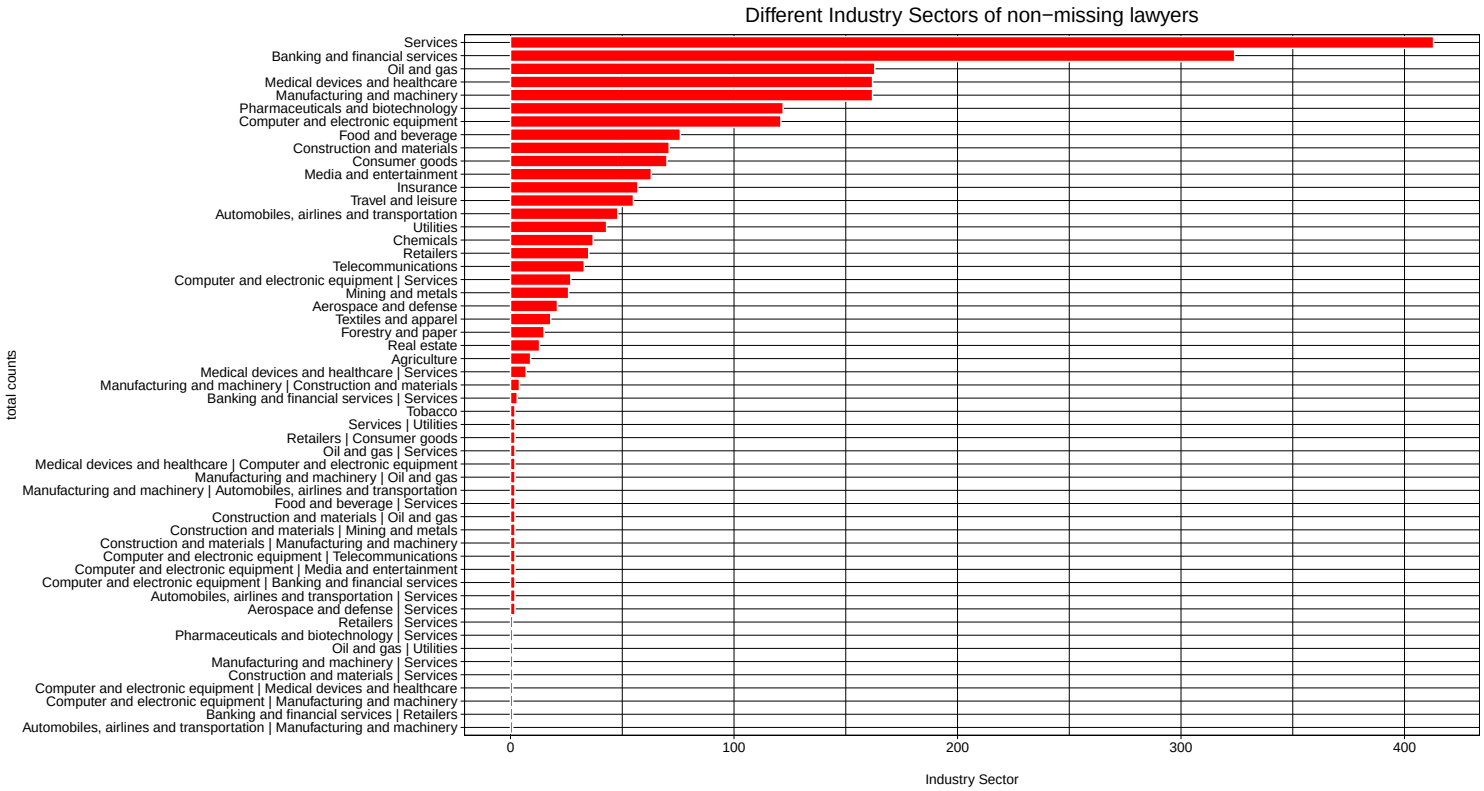
Distribution of the missing and non-missing lawyers

What is the distribution of the missing lawyers by deal value (category), year, and industry sector? Histogram for each plus mean, standard, median.

Counts of Missing Lawyers



Counts of non-missing lawyers



Chi-Square Test of Independence

The top 15 most industry sector among missing lawyer deals

```
## Selecting by n
```

```
> contingency %>%
+     kbl(caption = "Contingency Table of missing and non-missing lawyer deals by 15 Industry",
+         booktabs = T) %>%
+     kable_styling(latex_options = c("striped", "hold_position"))
```

Table 1: Contingency Table of missing and non-missing lawyer deals by 15 Industry

	missing	non_missing
Oil and gas	245	163
Services	157	413
Banking and financial services	111	324
Medical devices and healthcare	74	162
Manufacturing and machinery	57	162
Food and beverage	49	76
Pharmaceuticals and biotechnology	48	122
Computer and electronic equipment	44	121
Utilities	42	43
Construction and materials	37	71
Media and entertainment	32	63
Automobiles, airlines and transportation	28	48
Consumer goods	27	70
Insurance	20	57
Chemicals	18	37

- Blue color indicates that the observed value is higher than the expected value if the data were random
- Red color specifies that the observed value is lower than the expected value if the data were random

From this plot generated by `Pearson's chi-squared`, we can tell that `oil and gas` has more missing lawyers than expected while `oil and gas` has much lower deals of non-missing lawyers than expected.

```
> chisq <- chisq.test(contingency)
> chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency
## X-squared = 176.33, df = 14, p-value < 2.2e-16
```

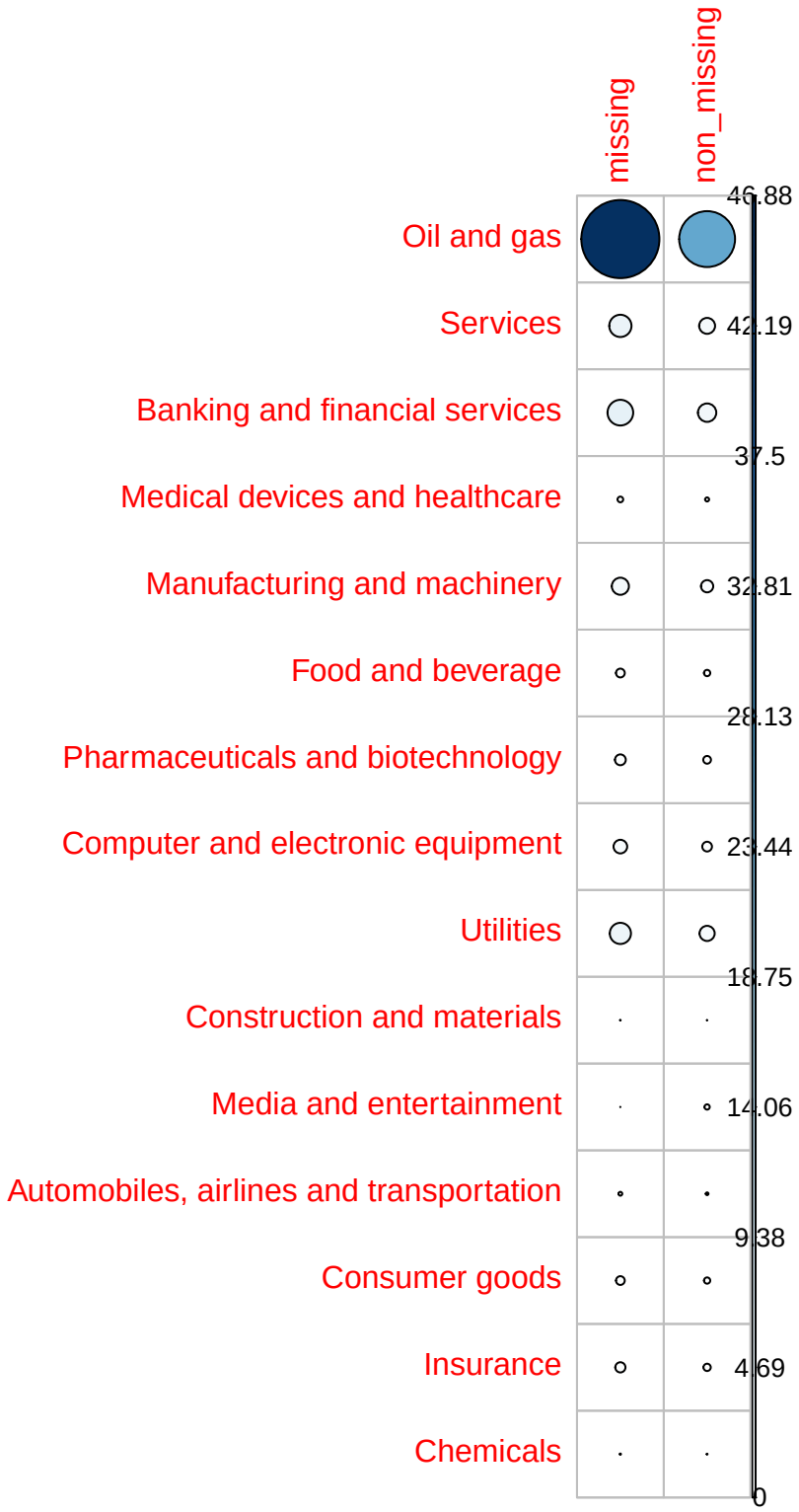
From this $p - value \approx 0$, we can tell that industry sector are missing versus non-missing are statistically significantly associated.

If we want to know the most contributing cells to the total Chi-square score, you just have to calculate the Chi-square statistic for each cell:

$$r = \frac{o - e}{\sqrt{e}}$$

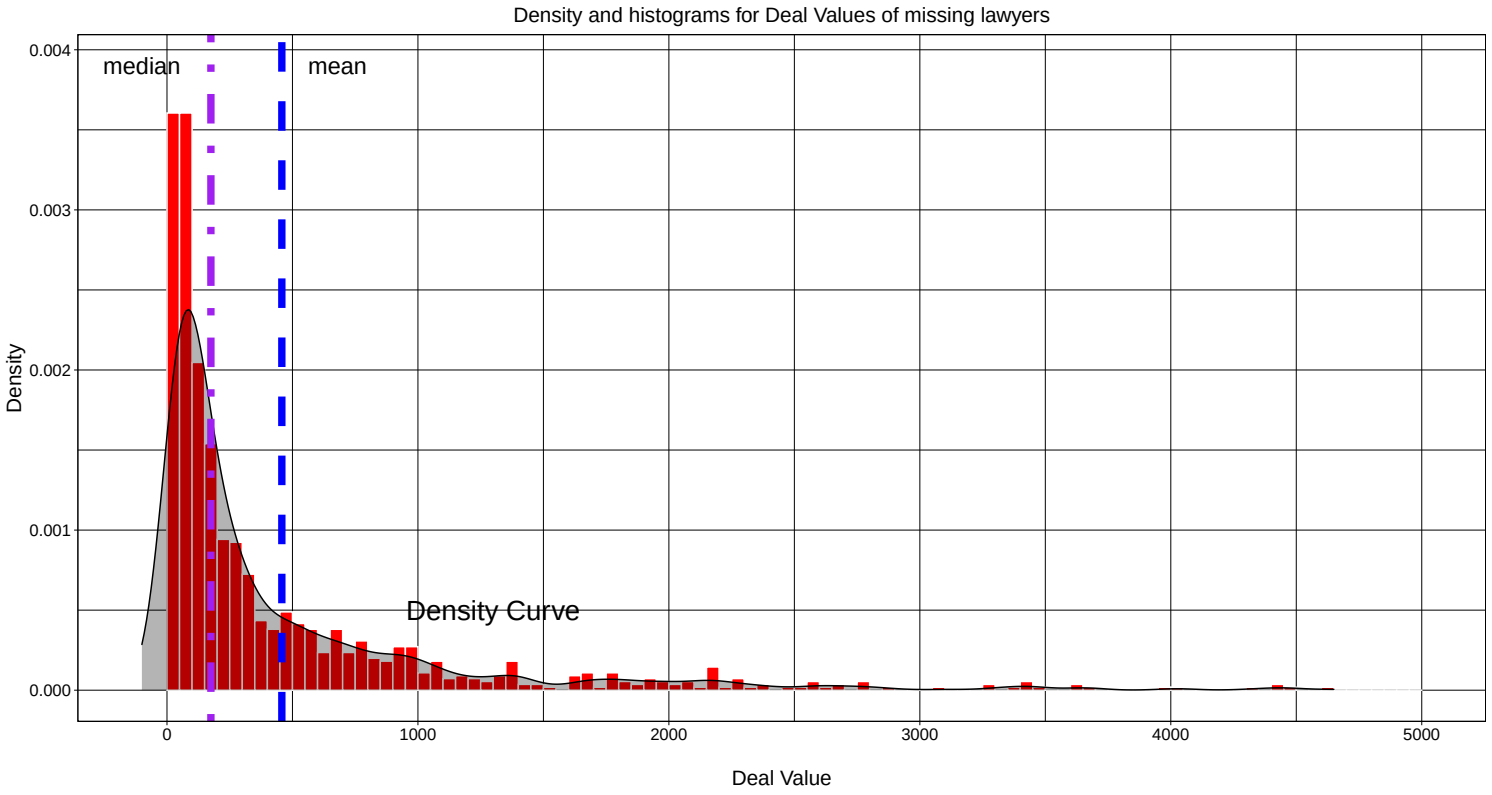
$$contrib = \frac{r^2}{\chi^2}$$

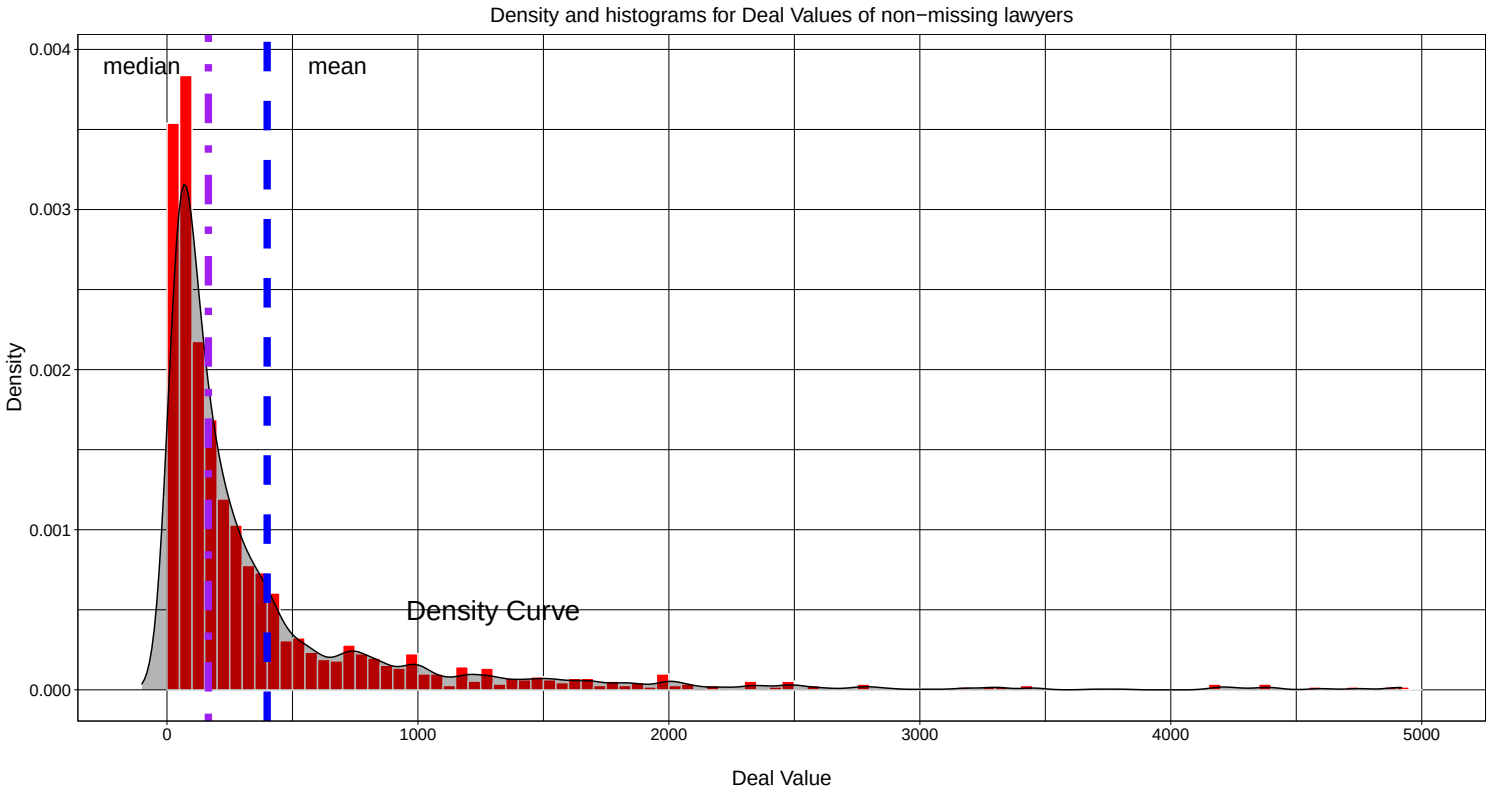
```
> # Visualize the contribution
> corrplot(contrib, is.corr = FALSE, outline = T)
```



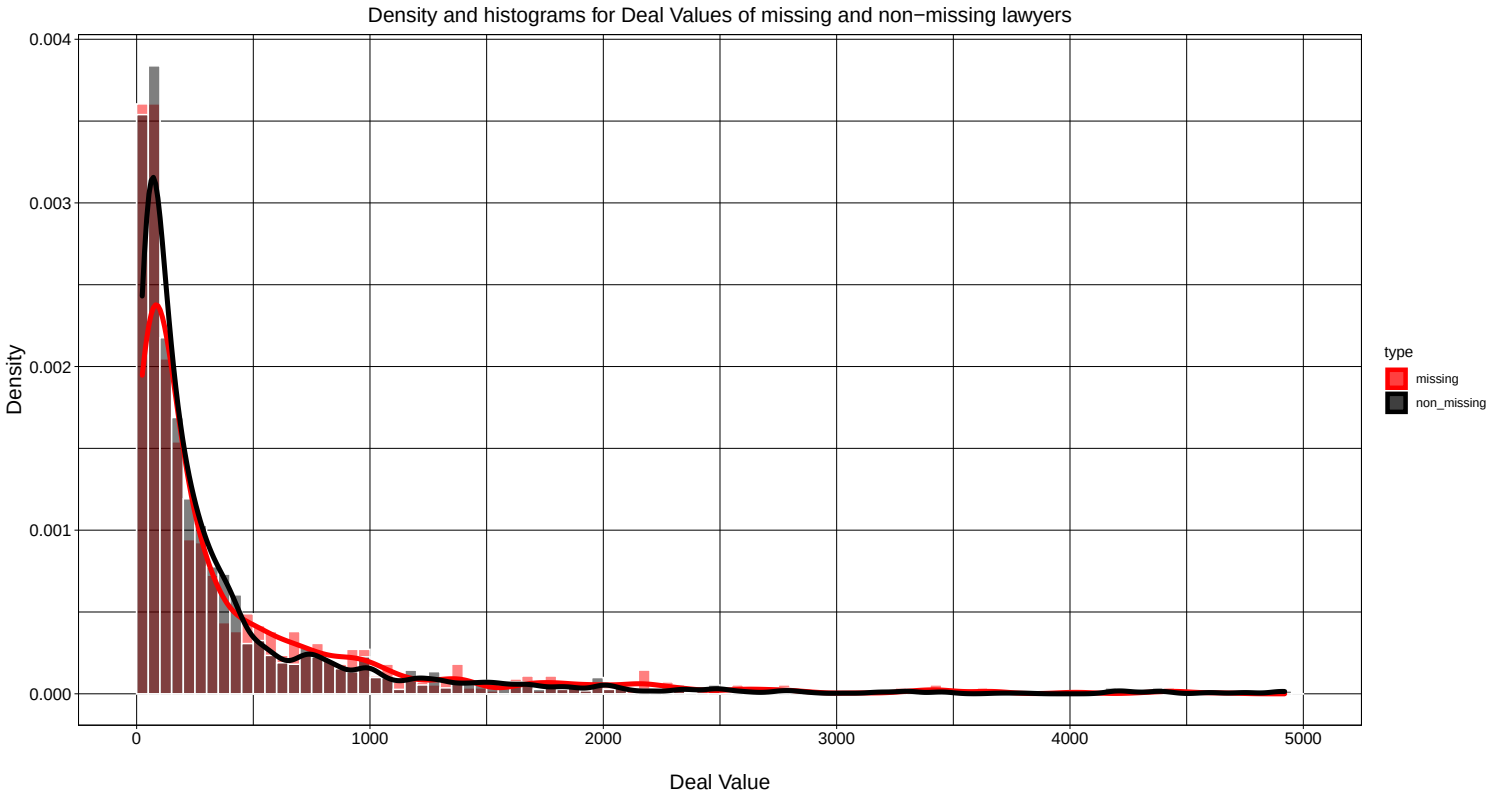
From this percentage contribution correlation plot, we can tell that `oil and gas` makes huge statistically significant to chi-squared values, which means that `oil and gas` industry sector is a main factor to determine missing versus non-missing lawyer in deals.

Distribution and Density curve of deal values of missing lawyers





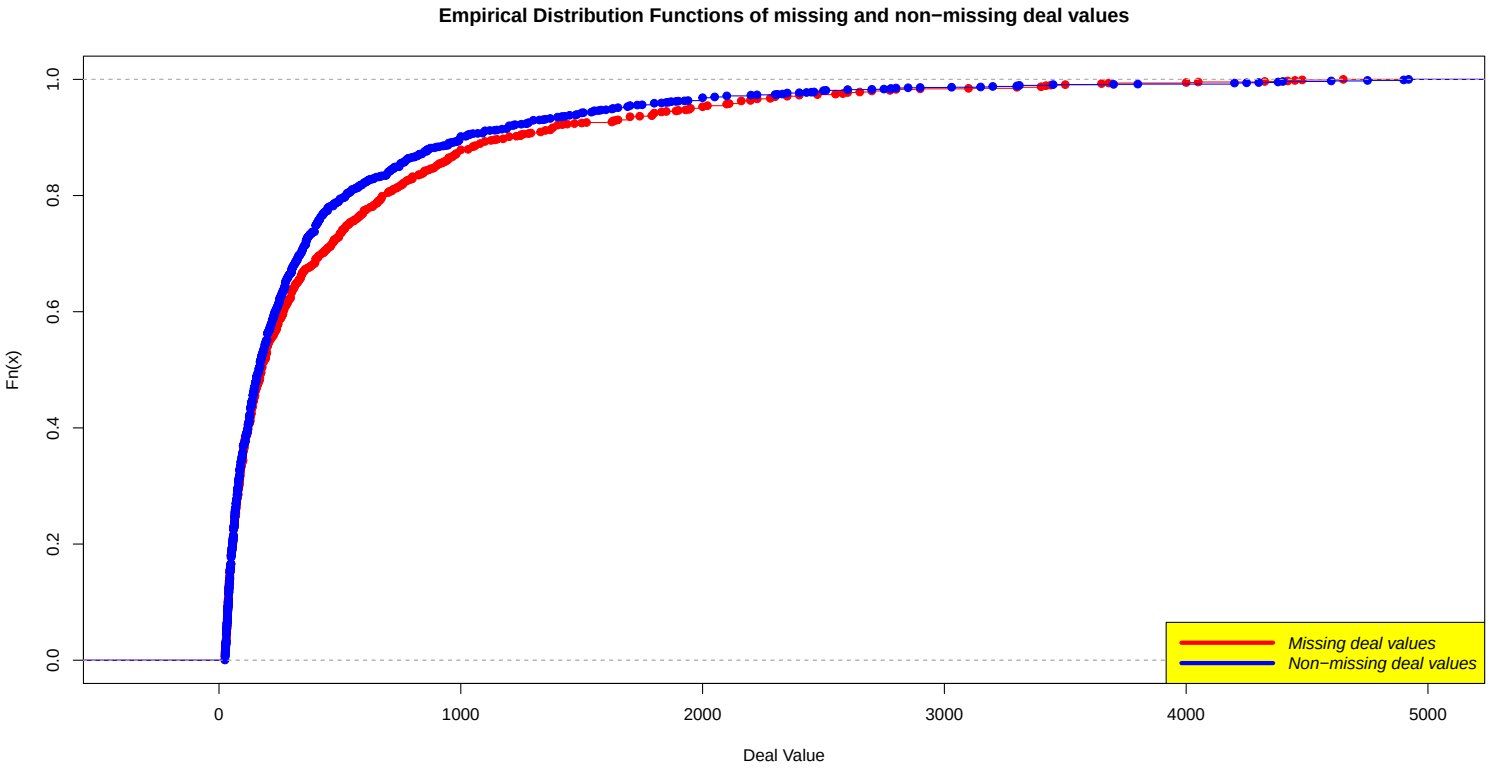
Put them together to compare



Two-sample Kolmogorov–Smirnov test

Empirical distributions of deal values for missing and non-missing lawyer deals

```
> plot(ecdf(dat %>%
+     filter(type == "missing") %>%
+     pull(value1)), col = "red", main = "Empirical Distribution Functions of missing and non-missing deal value
+     xlab = "Deal Value")
> lines(ecdf(dat %>%
+     filter(type == "non_missing") %>%
+     pull(value1)), col = "blue")
> legend(x = "bottomright", legend = c("Missing deal values", "Non-missing deal values"),
+     col = c("red", "blue"), lty = c(1, 1), lwd = 4, bg = "yellow", seg.len = 6, text.font = 3)
```



Two-sample Kolmogorov–Smirnov test

The empirical distribution function F_n for n independent and identically distributed (i.i.d.) ordered observations X_i is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty,x]}(X_i)$$

where $I_{[-\infty,x]}(X_i)$ is the indicator function, equal to 1 if $X_i \leq x$ and equal to 0 otherwise.

The Kolmogorov–Smirnov test may also be used to test whether two underlying one-dimensional probability distributions differ. In this case, the Kolmogorov–Smirnov statistic is

$$D_{n,m} = \sup_x \mid F_{1,n}(x) - F_{2,m}(x)$$

where $F_{1,n}$ and $F_{2,m}$ are the empirical distribution functions of the first and the second sample respectively, and sup is the supremum function.

```
> miss <- dat %>%
+     filter(type == "missing") %>%
+     pull(value1)
> non_miss <- dat %>%
+     filter(type == "non_missing") %>%
+     pull(value1)
> ks.test(miss, non_miss)
```

```
## Warning in ks.test(miss, non_miss): p-value will be approximate in the presence
## of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  miss and non_miss
## D = 0.069539, p-value = 0.001608
## alternative hypothesis: two-sided
```

From this p-value, 0.001608, we can say that missing and non-missing lawyer deal values are from different distribution since the null hypothesis is two samples are drawn from the same distribution. However, this p-value is not too small and from the ECDF plot, they are quite close, in light of p-value isn't always reliable, so I think the distribution of missing lawyer deal values are almost very close that of non-missing counterpart.

Mean, Median and Standard Deviation of deal values

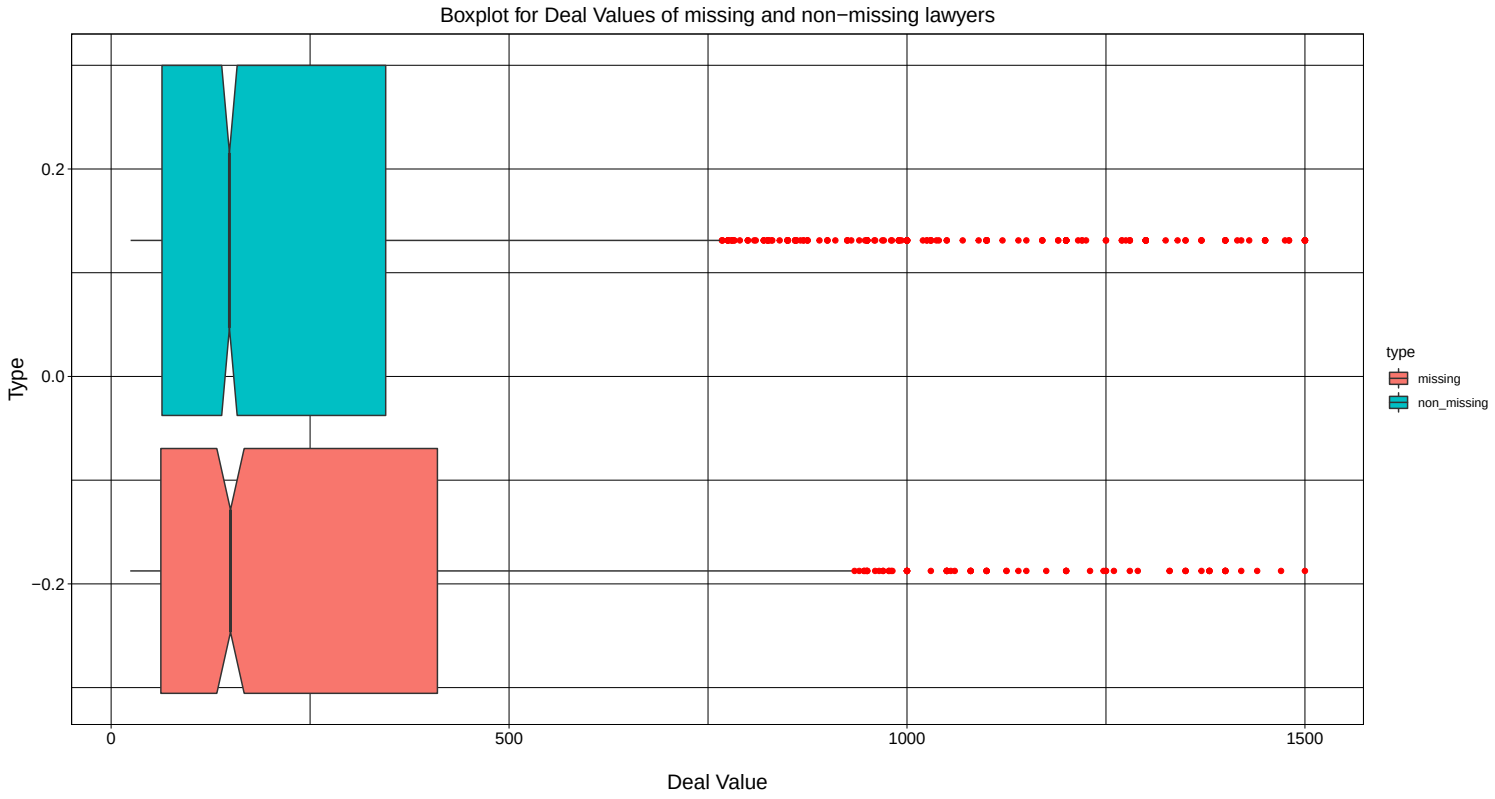
```
> miss_non_miss %>%
+     group_by(type) %>%
+     summarise(`:=`(min, min(value1, na.rm = T)), `:=`(mean, mean(value1, na.rm = T)),
+         `:=`(median, median(value1, na.rm = T)), `:=`(max, max(value1, na.rm = T)),
+         `:=`(sd, sd(value1, na.rm = T))) %>%
+     kbl(caption = "Summary Stats Table comparing missing and non-missing", booktabs = T) %>%
+     kable_styling(latex_options = c("striped", "hold_position"))
```

Table 2: Summary Stats Table comparing missing and non-missing

type	min	mean	median	max	sd
missing	24.0	687.6272	180	39000	2099.451
non_missing	24.5	494.4269	169	32700	1308.376

Boxplot for missing and non-missing

```
> miss_non_miss %>%
+     filter(value1 <= 1500) %>%
+     ggplot(aes(x = value1)) + geom_boxplot(aes(fill = type), na.rm = F, notch = T,
+     varwidth = T, orientation = "y", outlier.colour = "red") + theme_linedraw() +
+     labs(x = "\nDeal Value", y = "Type", title = "Boxplot for Deal Values of missing and non-missing lawyers")
+     theme(axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12),
+         axis.title = element_text(size = 15), plot.title = element_text(hjust = 0.5,
+             size = 15))
```

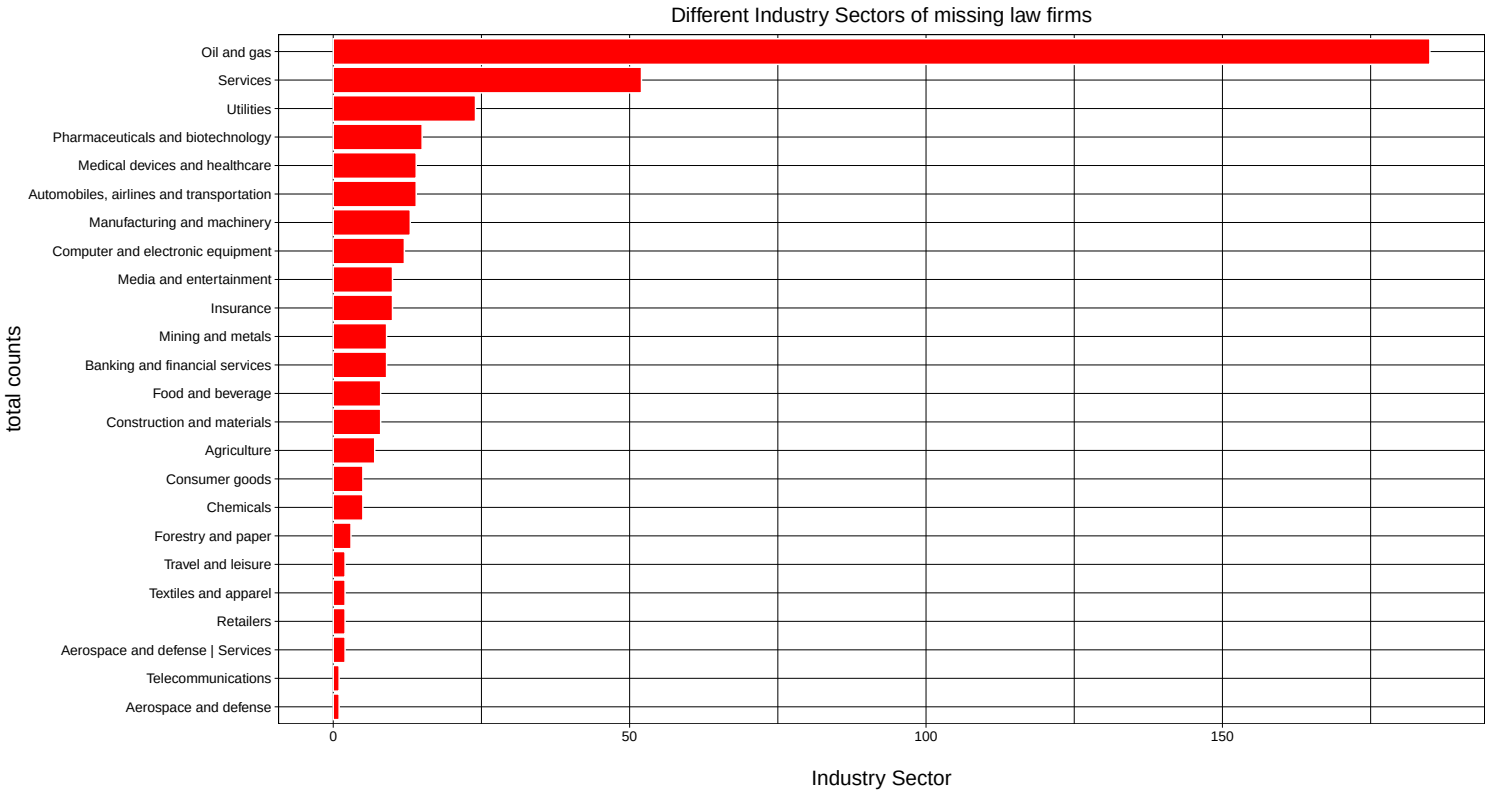


How many law firms are unknown (NA and “not disclosed”)?

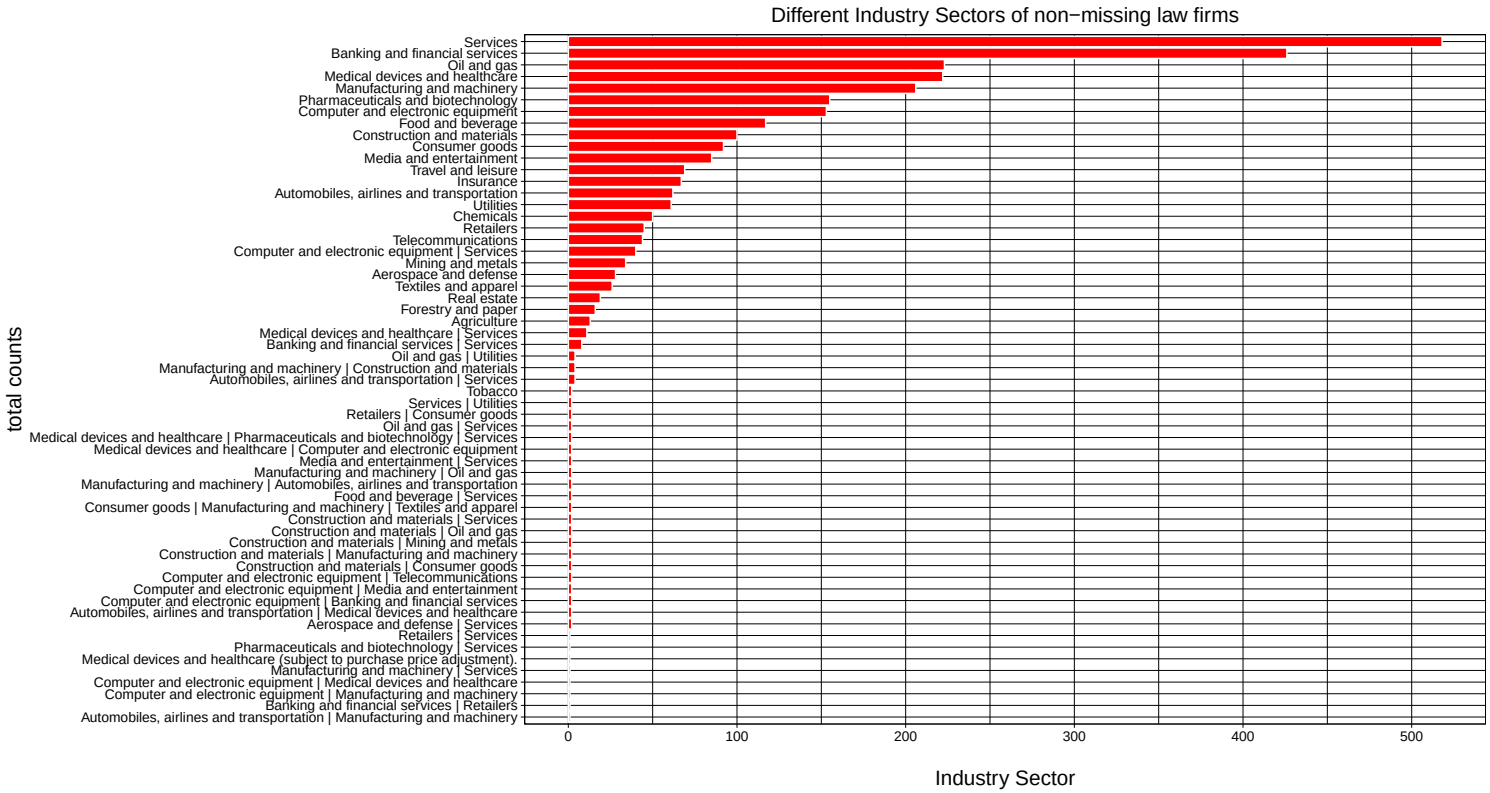
What is the distribution of the missing law firms by the size (in dollars) of the deal, year, and industry? Histogram each.

Extract corresponding dataset

Bar plots for missing law firms



Bar plots for non_missing law firms

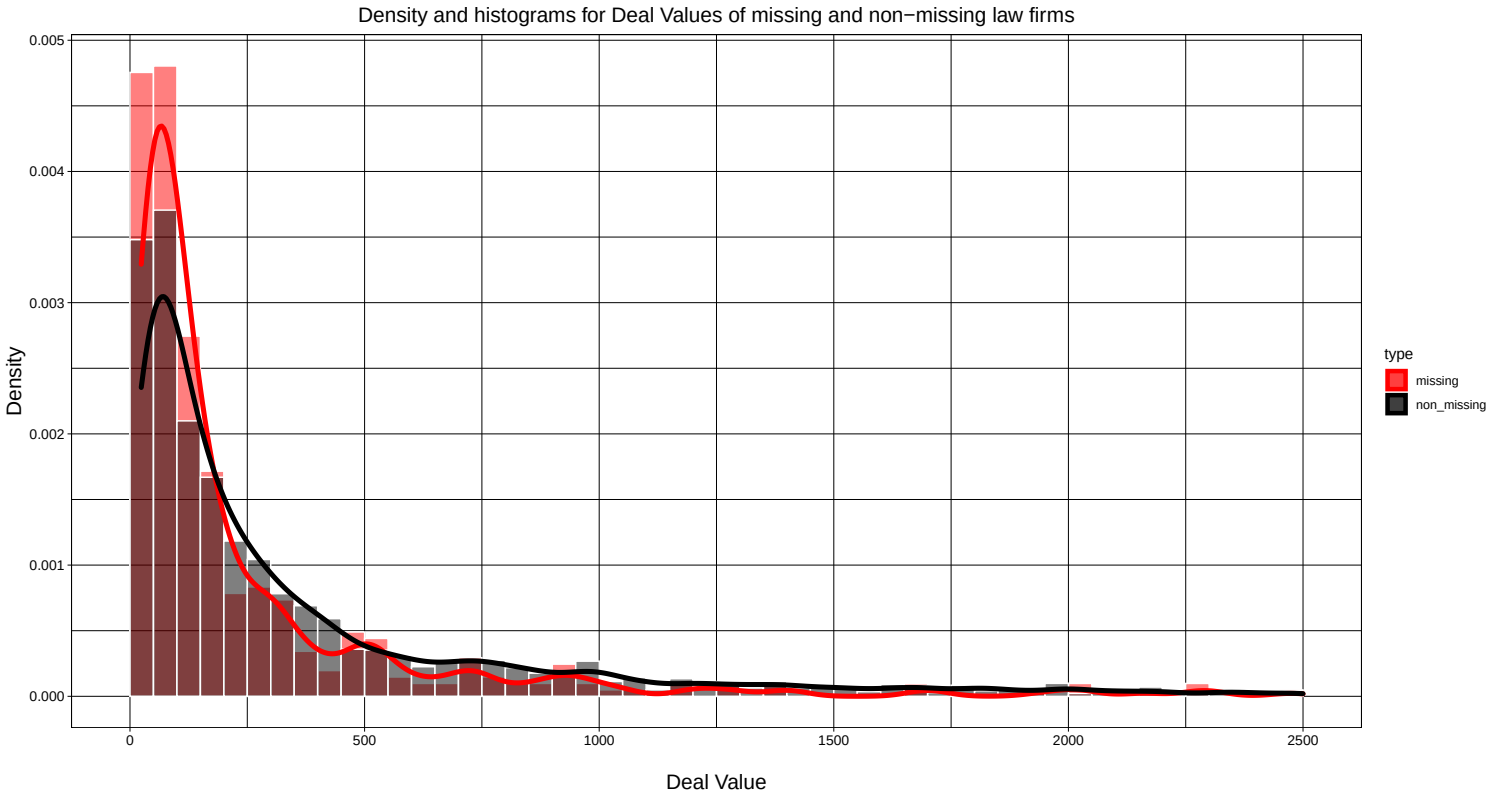


Histogram and density curves to compare them

```
> missing_law_firm <- missing_law_firm %>%
+     mutate(`:=`(type, "missing"))
> non_missing_law_firm <- non_missing_law_firm %>%
+     mutate(`:=`(type, "non_missing"))
> miss_non_miss_law_firm <- rbind(missing_law_firm, non_missing_law_firm)
> View(miss_non_miss_law_firm)
```

Combine two dataset marked by types

Plot it based on types



Mean, Median and Standard Deviation

```
> miss_non_miss_law_firm %>%
+     group_by(type) %>%
+     summarise(`:=`(min, min(value1, na.rm = T)), `:=`(mean, mean(value1, na.rm = T)),
+         `:=`(median, median(value1, na.rm = T)), `:=`(max, max(value1, na.rm = T)),
+         `:=`(sd, sd(value1, na.rm = T))) %>%
+     kbl(caption = "Summary Stats Table comparing missing and non-missing", booktabs = T) %>%
+     kable_styling(latex_options = c("striped", "hold_position"))
```

Table 3: Summary Stats Table comparing missing and non-missing

type	min	mean	median	max	sd
missing	25	332.3452	107	9730	882.8552
non_missing	24	590.8620	185	39000	1694.6415

How many deal attorneys are missing biographical information?

What is the distribution of the missing data – are these mostly from earlier years, for example?

Load the no-matched data set

```
> com_deal_lawyer_firm_last_no_match <- read_csv("../data/confidence_match/com_deal_lawyer_firm_last_no_match.cs
> # View(com_deal_lawyer_firm_last_no_match)
```

attorneys without biographical info

```
> att_miss_bio <- com_deal_lawyer_firm_last_no_match %>%
+     drop_na(First) %>%
+     mutate(`:=`(year, str_match(string = `Signing date`, pattern = "\\d\\d\\d\\d")))
> # View(att_miss_bio)
```

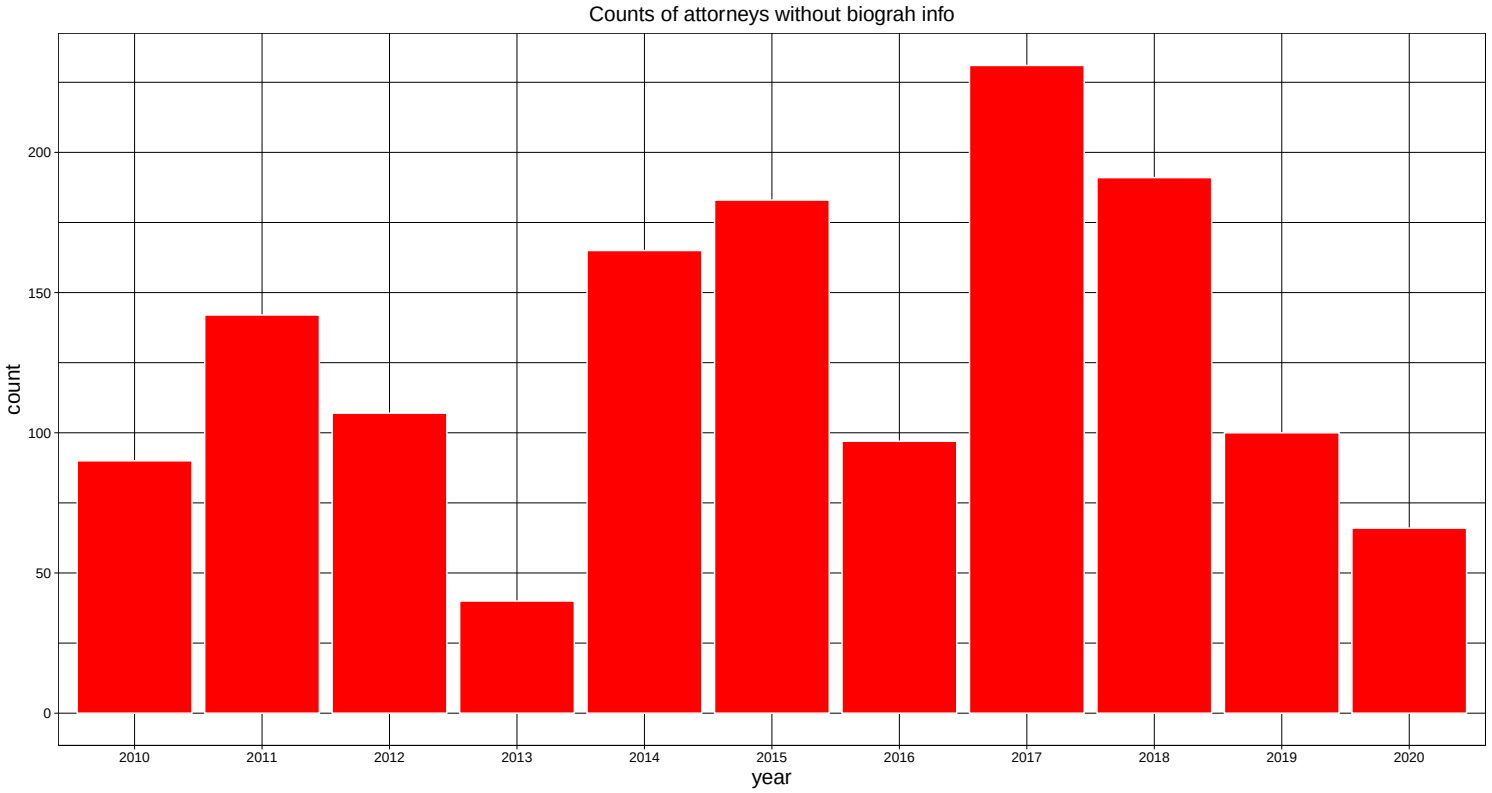
```
>
> att_miss_bio %>%
+     distinct(First, Last) %>%
+     nrow
```

```
## [1] 904
```

There are 904 attorneys without biographical info.

Distribution of Years

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Distribution of deal type

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

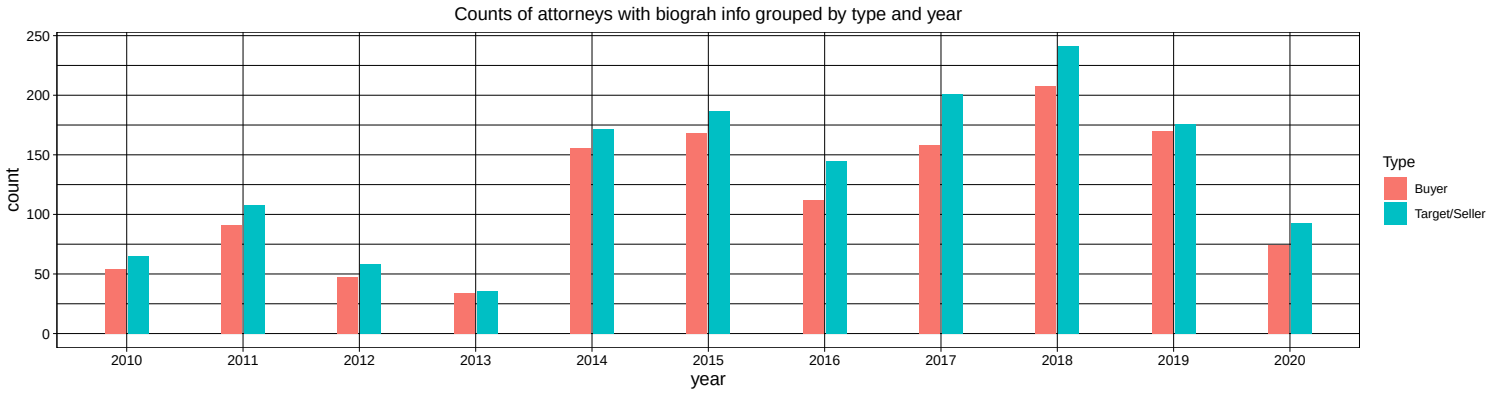
Distribution of attorneys with biograh info

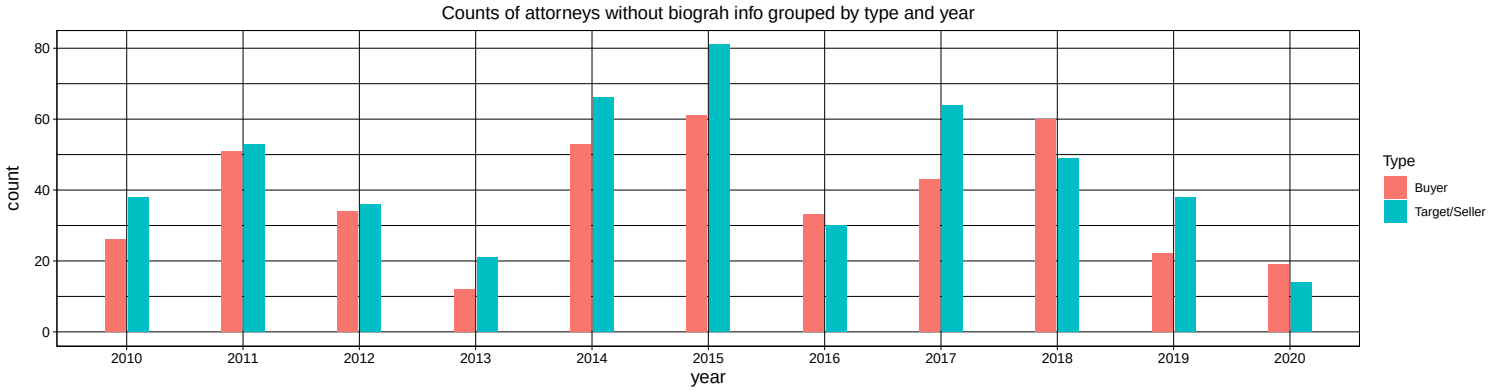
```
> three_matched_stacked <- read_csv("../data/confidence_match/three_matched_stacked.csv",
+     col_types = cols(.default = "c"))
> # View(three_matched_stacked)
>
> # If duplicates, keep attorneys from MA only
> dis_three_matched_stacked <- three_matched_stacked %>%
+     distinct(Deal_number, `Signing date`, First, Last, Law_Firm, .keep_all = T)
> # View(dis_three_matched_stacked)
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

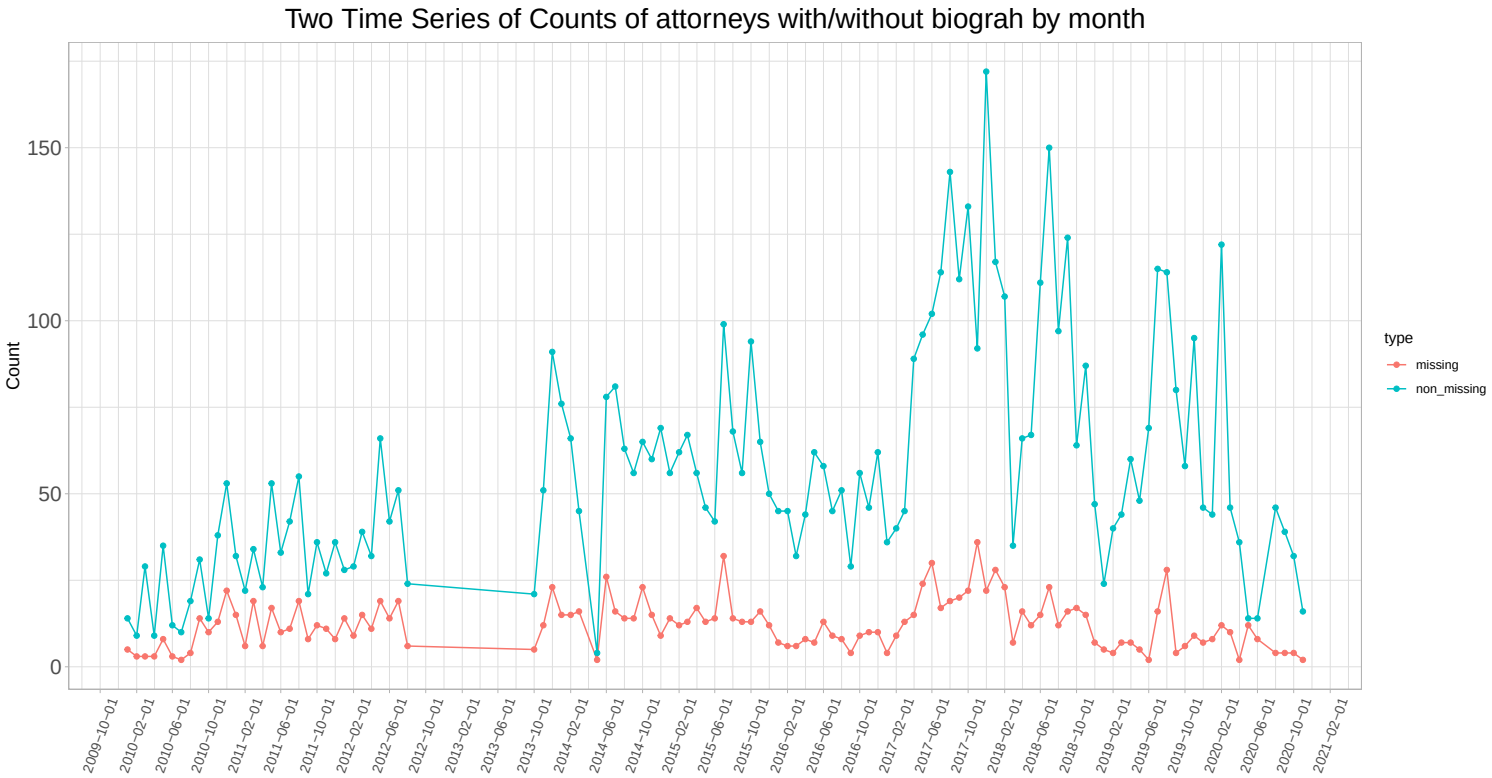
Compare them together

```
> p_with/p_without
```





Two time series by months of counts of attorneys with and without biograh info



From these two time series by month, we can tell numbers of attorneys with/without biograh info have very close trend but absolute number. The missing number is less than non-missing number, but they have the same trend.

How many deals are missing attorneys' biographical information?

```
> att_miss_bio %>%
+     distinct(Deal_number) %>%
+     nrow()
```

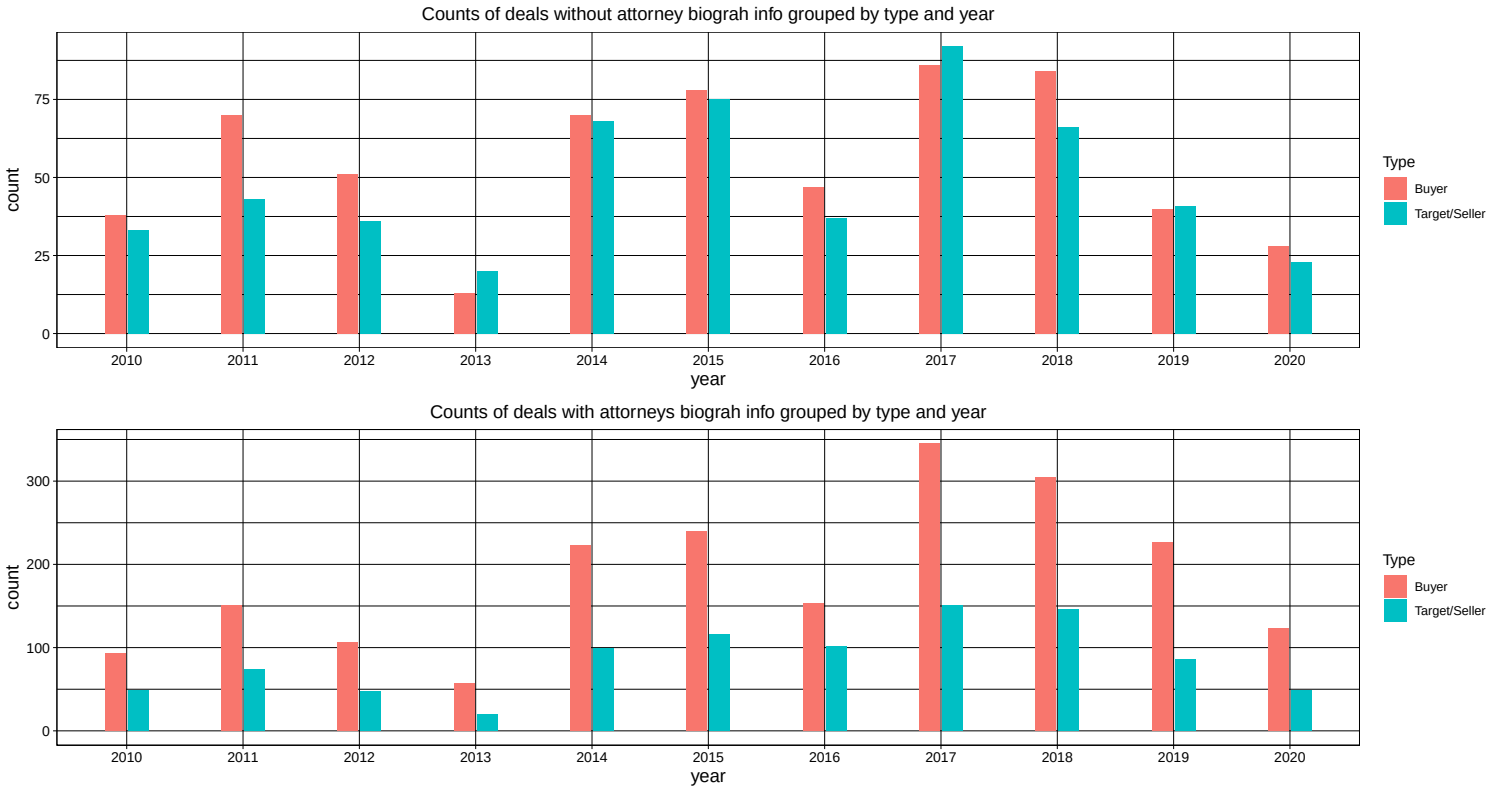
```
## [1] 1139
```

There are 1139 deals without attorneys' biograh info.

Counts of deals without attorneys' biograh info grouped by Year and Deal Type.

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Create some summary tables

Create a summary table of the number of deals, gender breakdown, age distribution (of lawyers appearing on the deal); school distribution (i.e., most common law school attended), law firm distribution (i.e., most common law firms appearing on the deal); types of deals (e.g., health care; financial services, etc.); size of deal

Number of deals, gender, age

a summary table of the number of deals, gender breakdown, age distribution (of lawyers appearing on the deal)

We consider 25 years old as average ages when students graduate from law school.

```
> dis_three_matched_stacked %>%
+     drop_na(Gender, age_breaks) %>%
+     group_by(Gender, age_breaks) %>%
+     count() %>%
+     arrange(desc(n)) %>%
+     kbl(caption = "Summary Stats Table comparing missing and non-missing", booktabs = T) %>%
+     kable_styling(latex_options = c("striped", "hold_position"))
```

Table 4: Summary Stats Table comparing missing and non-missing

Gender	age_breaks	n
Male	45 < age <= 60	2713
Male	30 < age <= 45	1027
Male	age > 60	1011
Female	45 < age <= 60	295
Female	30 < age <= 45	213
Female	age > 60	68
Male	age <= 30	15
Female	age <= 30	3

school and law distribution

school distribution (i.e., most common law school attended), law firm distribution (i.e., most common law firms appearing on the deal)

Please see attachment csv file for full list data here.

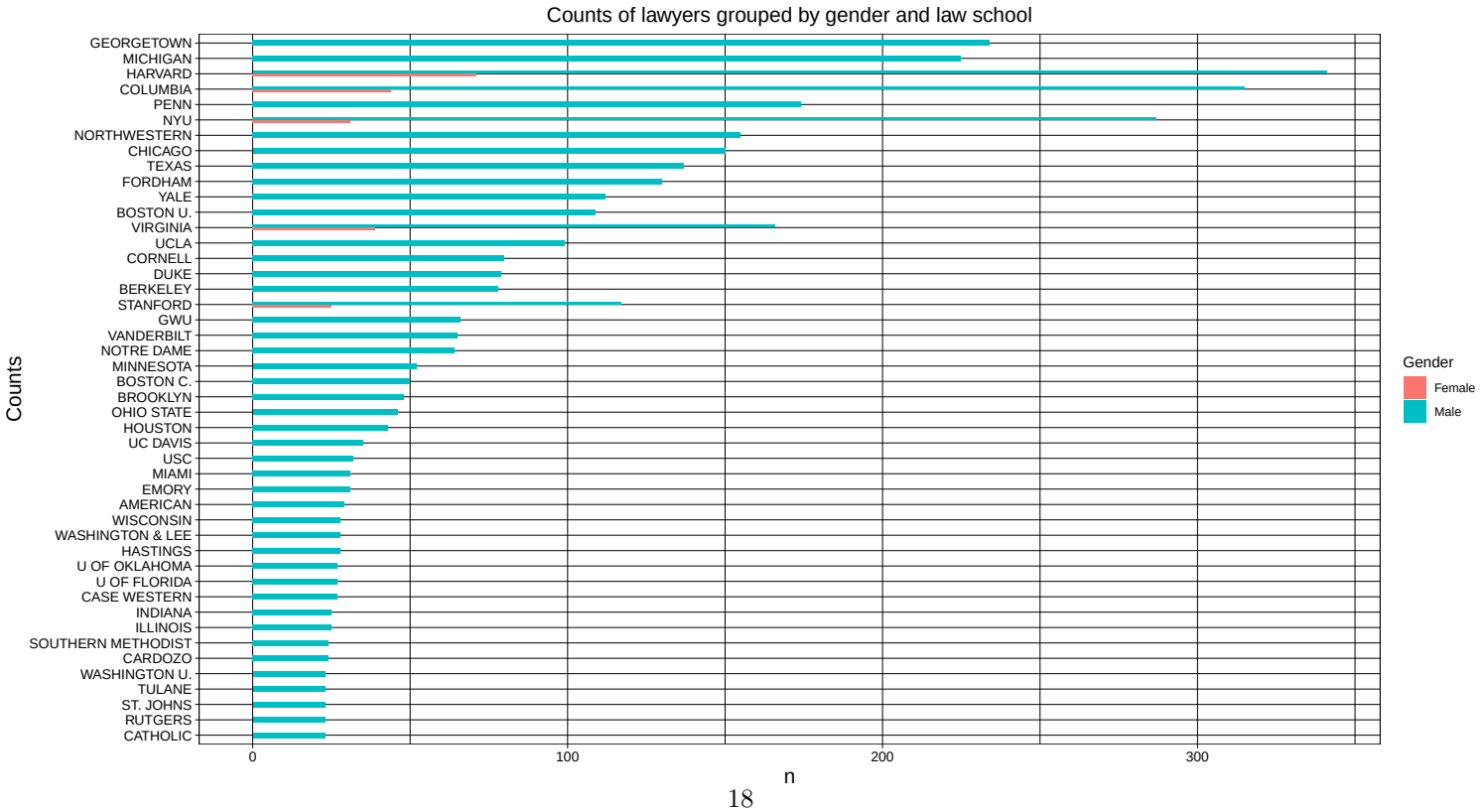
```
> dis_three_matched_stacked %>%
+     drop_na(Gender, `Law School`) %>%
+     group_by(`Law School`, Law_Firm) %>%
+     count() %>%
+     arrange(desc(n)) %>%
+     write_csv(file = "../data/deal_lawyer/distri_law_sch_firm.csv")
```

```
> dis_three_matched_stacked %>%
+     drop_na(Gender, `Law School`) %>%
+     group_by(Gender, `Law School`) %>%
+     count() %>%
+     arrange(desc(n)) %>%
+     ungroup(Gender, `Law School`) %>%
+     top_n(50) %>%
+     ggplot(aes(y = reorder(`Law School`, n), x = n)) + geom_col(position = position_dodge2(),
+     width = 0.4, orientation = "y", aes(fill = Gender)) + labs(title = "Counts of lawyers grouped by gender ar
+     y = "Counts") + theme_linedraw() + theme(axis.text.x = element_text(size = 10),
+     axis.text.y = element_text(size = 10), axis.title = element_text(size = 15),
+     plot.title = element_text(hjust = 0.5, size = 15))
```

```
## Selecting by n
```

Table 5: Top 30 most law school and law firm attended

Law School	Law_Firm	n
MICHIGAN	Kirkland & Ellis LLP	55
NORTHWESTERN	Kirkland & Ellis LLP	53
HARVARD	Latham & Watkins LLP	42
GEORGETOWN	Skadden, Arps, Slate, Meagher & Flom LLP	39
COLUMBIA	Wachtell, Lipton, Rosen & Katz	35
YALE	Wachtell, Lipton, Rosen & Katz	34
CHICAGO	Kirkland & Ellis LLP	32
NYU	Latham & Watkins LLP	32
FORDHAM	Skadden, Arps, Slate, Meagher & Flom LLP	30
COLUMBIA	Paul, Weiss, Rifkind, Wharton & Garrison LLP	29
HARVARD	Wachtell, Lipton, Rosen & Katz	28
COLUMBIA	Latham & Watkins LLP	27
NYU	Skadden, Arps, Slate, Meagher & Flom LLP	26
TEXAS	Latham & Watkins LLP	25
HARVARD	Sidley Austin LLP	24
NYU	Wachtell, Lipton, Rosen & Katz	24
PENN	Wachtell, Lipton, Rosen & Katz	24
COLUMBIA	Sullivan & Cromwell LLP	23
NYU	Willkie Farr & Gallagher LLP	23
UCLA	Jones Day	22
COLUMBIA	Kirkland & Ellis LLP	21
U OF OKLAHOMA	Alston & Bird LLP	21
HARVARD	Kirkland & Ellis LLP	20
MICHIGAN	Latham & Watkins LLP	20
BOSTON U.	Ropes & Gray LLP	19
HARVARD	Ropes & Gray LLP	19
NYU	Kirkland & Ellis LLP	19
TEXAS	Vinson & Elkins LLP	19
BOSTON U.	Goodwin Procter LLP	18
MINNESOTA	Faegre Baker Daniels LLP	18



types of deals

types of deals (e.g., health care; financial services, etc.); size of deal

```
> type_value_deal %>%
+     group_by(`Industry sector`) %>%
+     count(sort = T) %>%
+     ungroup(`Industry sector`) %>%
+     filter(n >= 10) %>%
+     kbl(caption = "Industry Sector Count", booktabs = T) %>%
+     kable_styling(latex_options = c("striped", "hold_position"))
```

Table 6: Industry Sector Count

Industry sector	n
Services	1199
Banking and financial services	827
Medical devices and healthcare	531
Oil and gas	498
Manufacturing and machinery	455
Pharmaceuticals and biotechnology	389
Computer and electronic equipment	354
Food and beverage	282
Construction and materials	223
Consumer goods	203
Media and entertainment	181
Insurance	175
Travel and leisure	167
Utilities	135
Automobiles, airlines and transportation	121
Telecommunications	115
Retailers	108
Chemicals	99
Computer and electronic equipment | Services	99
Mining and metals	78
Aerospace and defense	72
Textiles and apparel	60
Forestry and paper	34
Real estate	33
Agriculture	31
Banking and financial services | Services	24
Medical devices and healthcare | Services	21
Manufacturing and machinery | Construction and materials	10
Services | Utilities	10

```
> type_value_deal %>%
+     drop_na(value1) %>%
+     mutate(`:=`(value_breaks, case_when(value1 <= 250 ~ "value <= 250", value1 >
+         250 & value1 <= 500 ~ "250 < value <= 500", value1 > 500 & value1 <= 1500 ~
+         "500 < value <= 1500", value1 > 1500 ~ "1500 < value"))) %>%
+     group_by(Type, value_breaks) %>%
+     count(sort = T) %>%
+     ungroup() %>%
+     kbl(caption = "Counts of deals grouped by types and value\\_breaks", booktabs = T) %>%
+     kable_styling(latex_options = c("striped", "hold_position"))
```

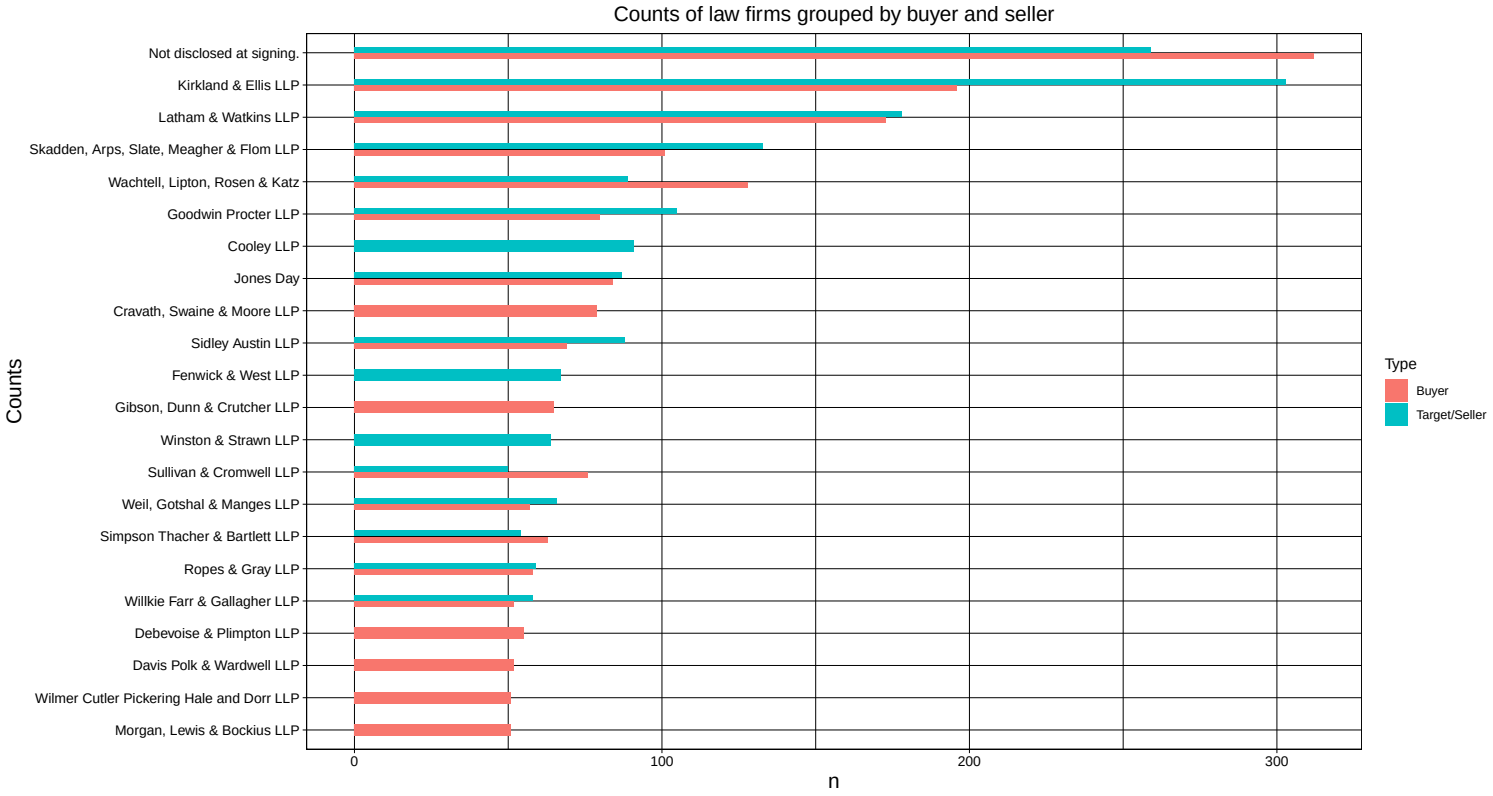
Table 7: Counts of deals grouped by types and value_breaks

Type	value_breaks	n
Target/Seller	value <= 250	1258
Buyer	value <= 250	1247
Target/Seller	250 < value <= 500	1140
Buyer	250 < value <= 500	1128
Target/Seller	1500 < value	975
Buyer	1500 < value	907

Buyer and seller of deals

Breakdown by buyer and seller in the deals. We think that law firms and M&A lawyers appear on both sides of the deal with roughly the same probability, but it would be helpful to know if this were true.

```
> encoded_merge_dl %>%
+     group_by(Law_Firm, Type) %>%
+     count(sort = T) %>%
+     ungroup(Law_Firm, Type) %>%
+     filter(n >= 50) %>%
+     ggplot(aes(y = reorder(Law_Firm, n), x = n)) + geom_col(position = position_dodge2(),
+     width = 0.4, orientation = "y", aes(fill = Type)) + labs(title = "Counts of law firms grouped by buyer and
+     y = "Counts") + theme_linedraw() + theme(axis.text.x = element_text(size = 10),
+     axis.text.y = element_text(size = 10), axis.title = element_text(size = 15),
+     plot.title = element_text(hjust = 0.5, size = 15))
```




```
> dis_three_matched_stacked %>%
+     filter(Source == "M&A") %>%
+     count(Type, sort = T) %>%
+     kbl(caption = "MA Lawyers grouped by Buyer/Seller", booktabs = T) %>%
+     kable_styling(latex_options = c("striped", "hold_position"))
```

Table 8: MA Lawyers grouped by Buyer/Seller

Type	n
Target/Seller	2685
Buyer	2660

Regression Analysis

Some basic regressions: e.g., regress gender on observable characteristics (e.g., firm, industry, size of deal). Explain what factors explain when women are more likely to appear on a deal.

First we used Multiple Logistic Regression with all variables

age coefficients

Law_school coefficient signs

Law_Firm coefficient signs

Industry Sector coefficient signs

Multiple Logistic Regression with all variables except for Law Schol and Law firm

Random Forest with the whole variables used

Variable Importance after fitting our model

Conclusion

From two method, Logistic Classification and Random Forest, we can tell that Age, Law_Firm, Law_school are obvious significant when predicting Gender. From huge positive coefficients bar plots, at least we can tell that Law_SchoolMeMPhis can make more likely Women on the deals. And age is another huge negative affects which can make huge Adverse conditions for women appearing on Deals. Deal Value and Type(buyer/Seller) aren't significant variables affecting women.

As for other law_school and law_firm coefficients, there are both positive and negative coefficients affecting women showing on Deals.

1. In total, there are 128 Law_school coefficients greater than 0, while 18 Law_school lower than 0, which means Law_school in general make positive affects on women showing on Deals.
2. However, there are 174 Law_Firm coefficients greater than 0, while 296 Law_Firm lower than 0, which means Law_Firm in general make negative affects on women showing on Deals.
3. There are 17 Industrial coefficients greater than 0, while 30 Industrial sector lower than 0, which means Industrial Sector in general make negative affects on women showing on Deals.