Compare missing versus non-missing Statistics*

Deal and Lawyer Studies

Li Yuan, li.yuan@vanderbilt.edu

27 July, 2021 00:36:15

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(janitor)
>
> library(patchwork)
> library(ggrepel)
> library(gghighlight)
> library(paletteer)
> library(ggExtra)
> library(ggbeeswarm)
> library(kableExtra)
> library(caret)
> library(randomForest)
> library(corrplot)
> library(lubridate)
> library(babynames)
```

Load data

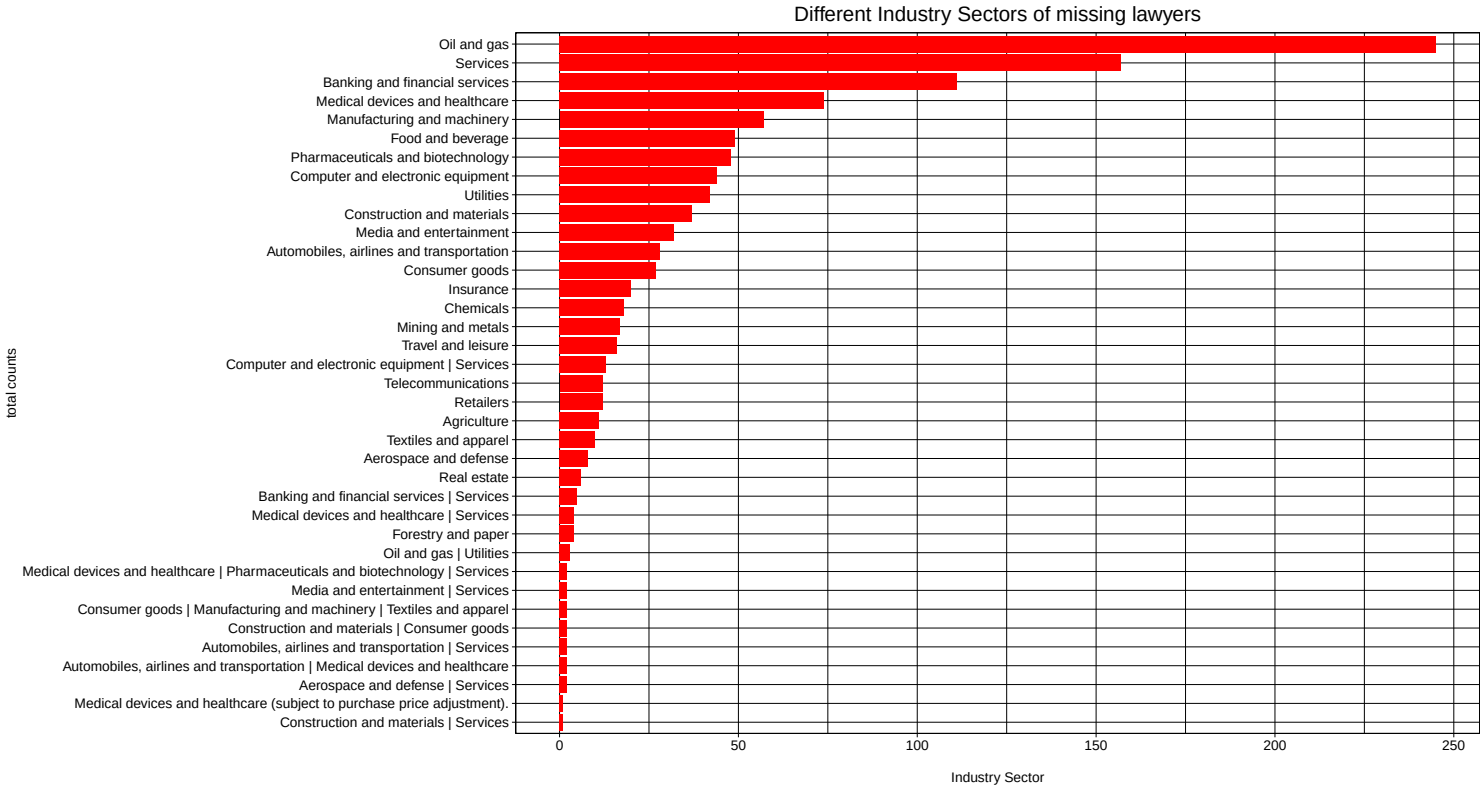
```
> deal <- read_csv("../data/deal/deal(1).csv", col_types = cols(.default = "c"))
> # View(deal)
>
> merge_deal_lawyer <- read_csv("../data/deal_lawyer/merge_deal_lawyer.csv")
> # View(merge_deal_lawyer)
>
> distin_com_lawyer <- read_csv("../data/lawyer/keep_MA_first.csv", col_types = cols(.default = "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")?

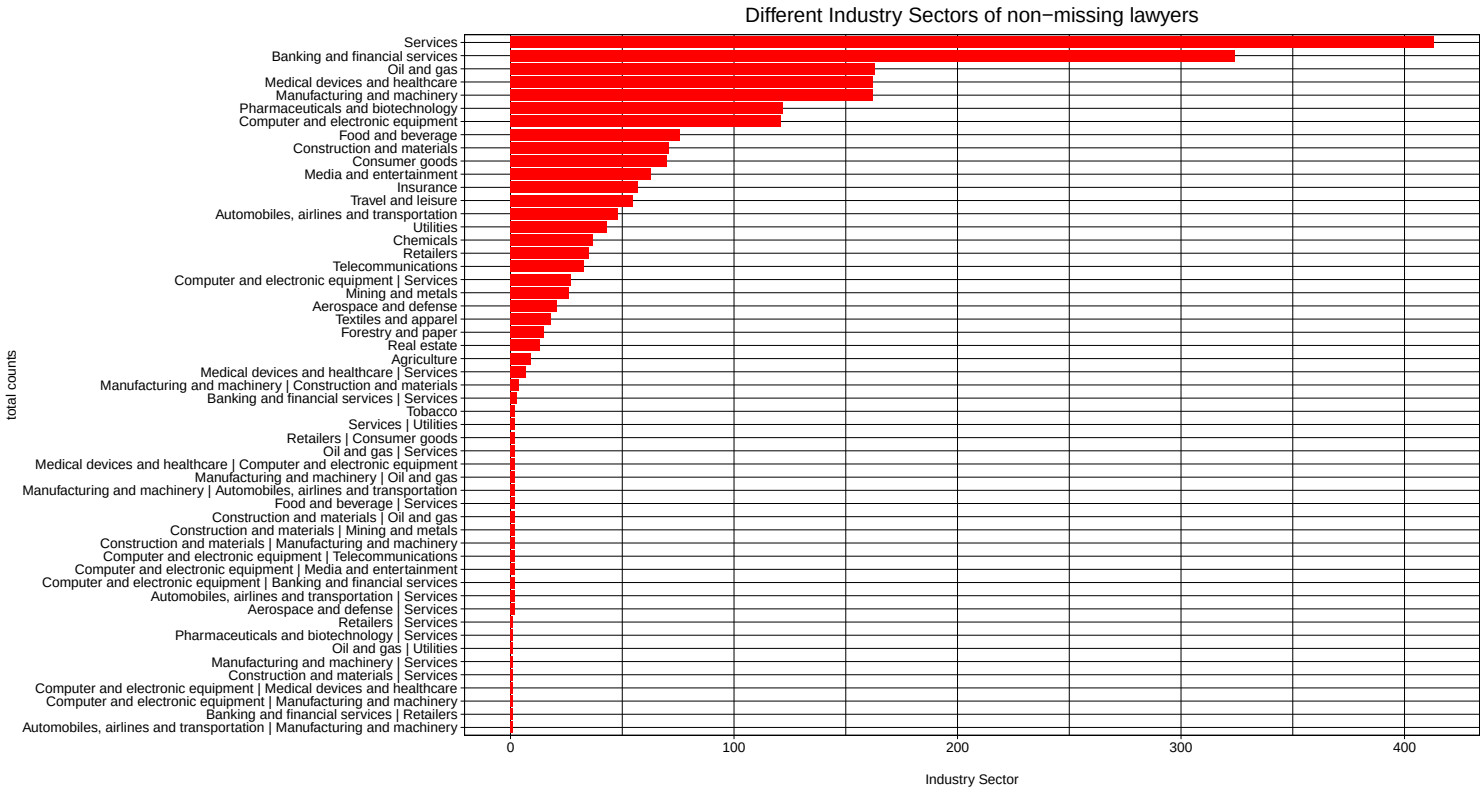
Distribution of the missing and non-missing lawyers

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
```

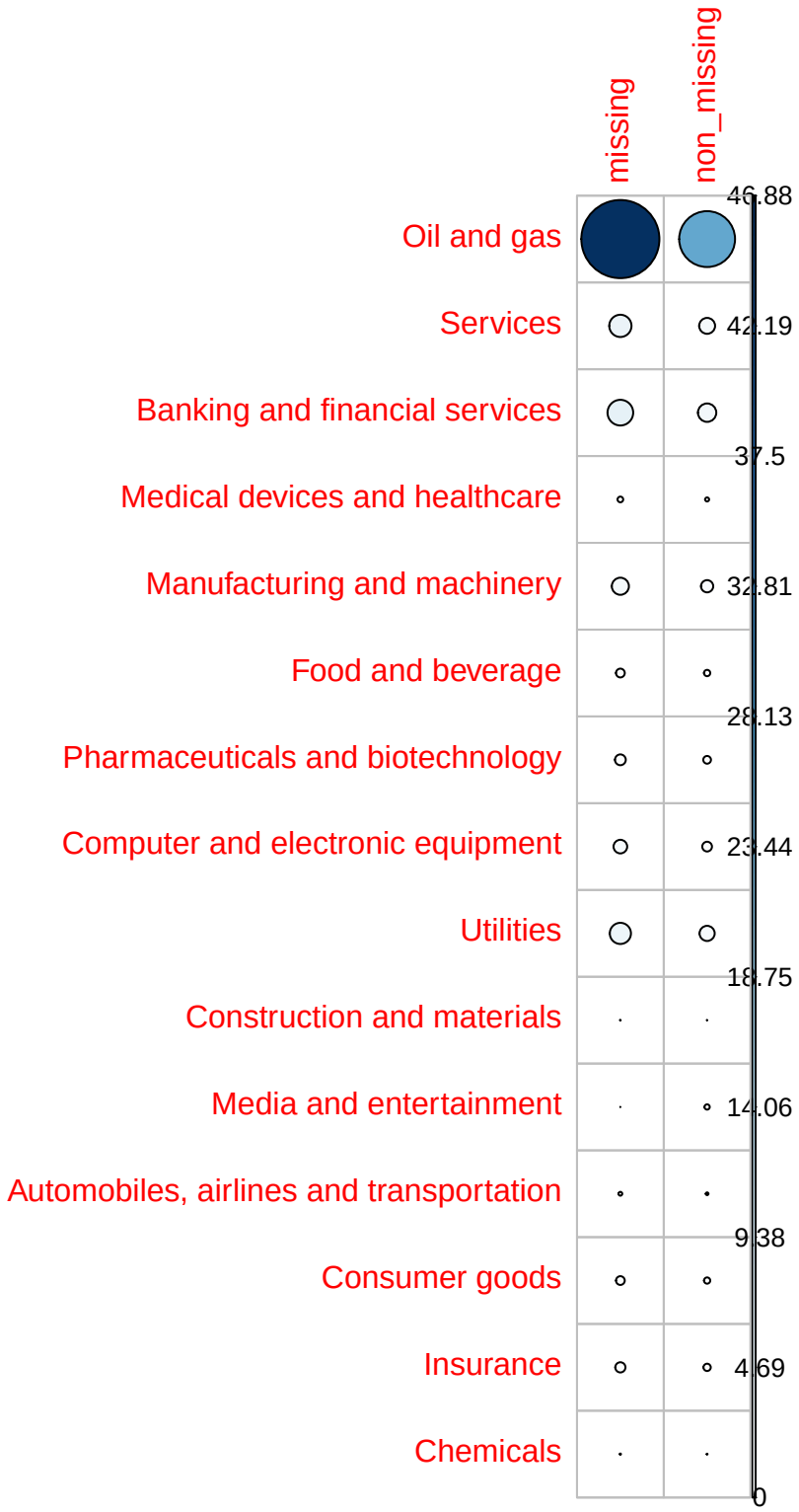
From this $p - value \approx 0$, we can tell that industry sector are missing versus non-missing are statistically significantly associated.

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:

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

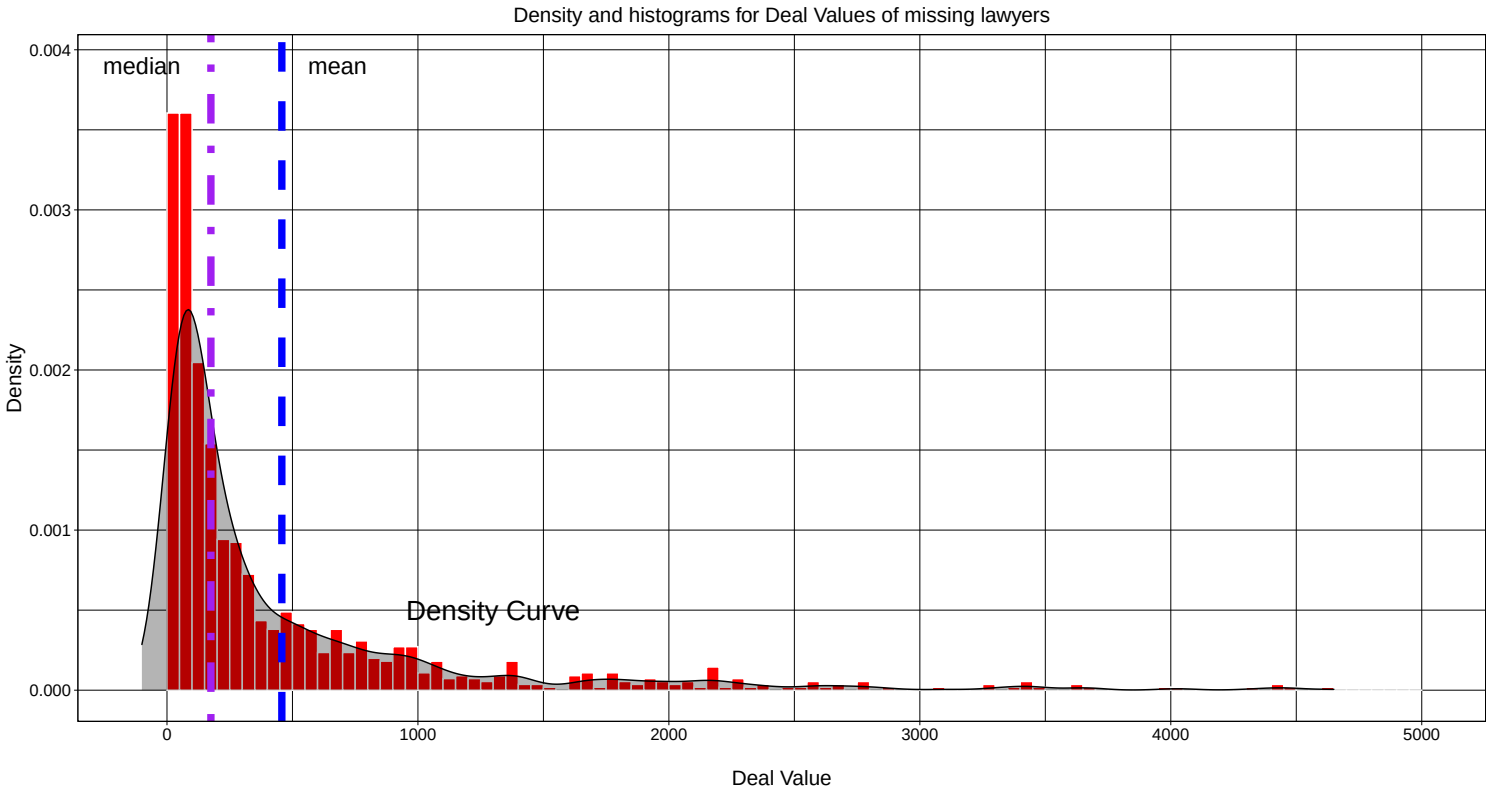
$$contrib = \frac{r^2}{\chi^2}$$

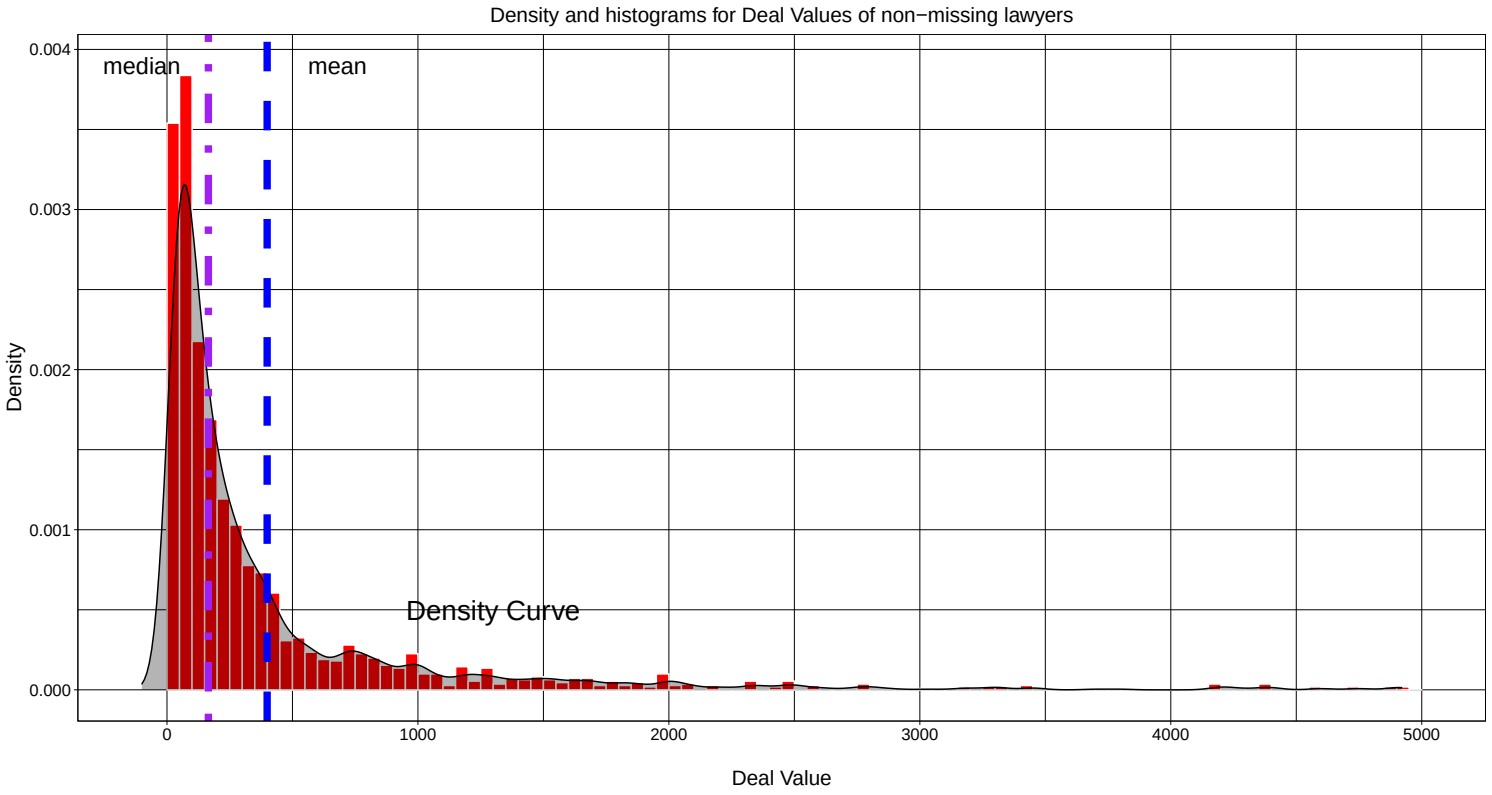
```
> # Visualize the contribution
> corrplot(contrib, is.corr = FALSE, outline = T)
```



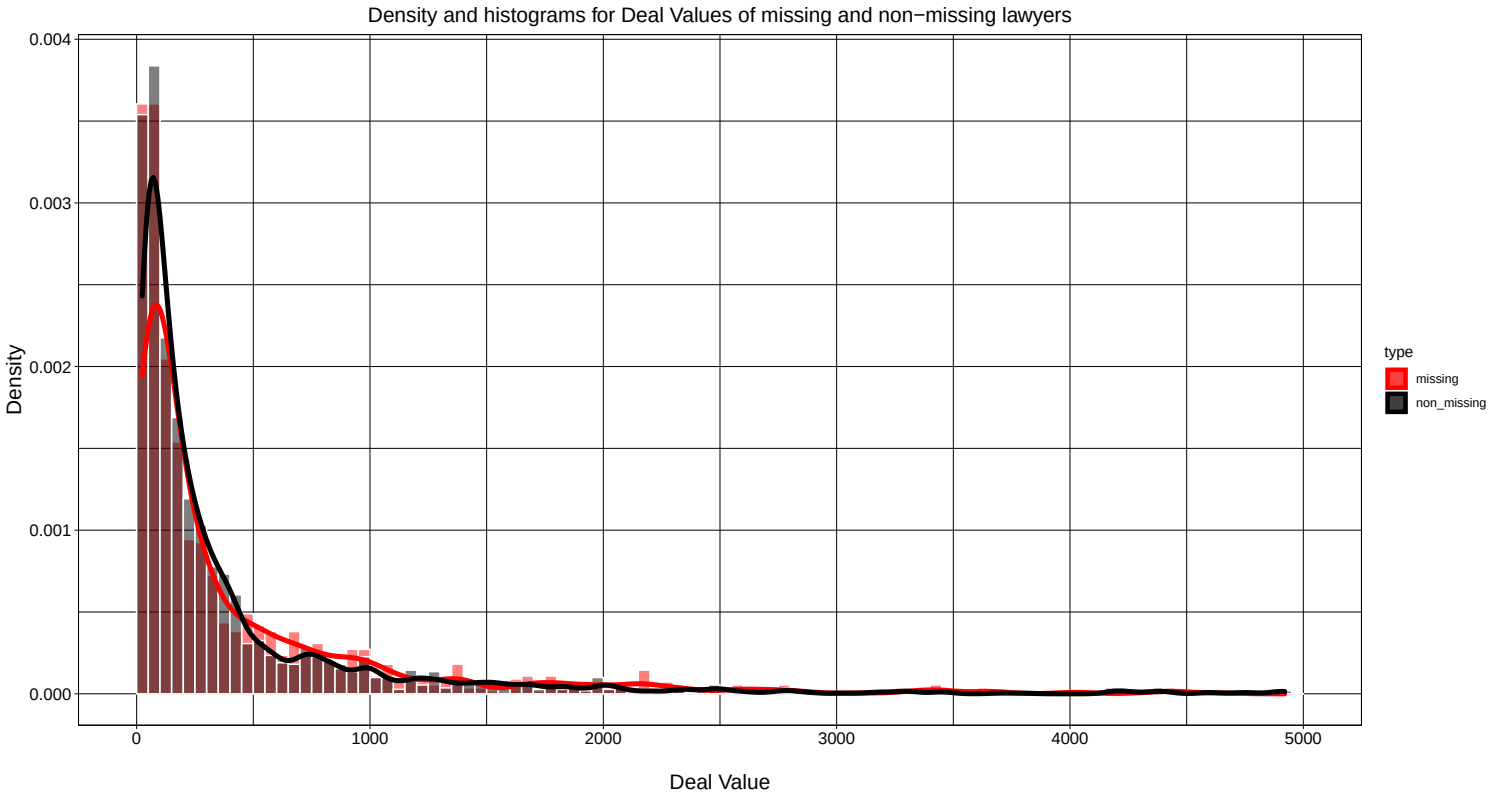
From this percentage contribution correlation plot, we can tell that `oil and gas` makes huge statistically significant to chi-squared values, which means that `oil and gas` industry sector is a main factor to determine missing versus non-missing lawyer in deals.

Distribution and Density curve of deal values of missing lawyers





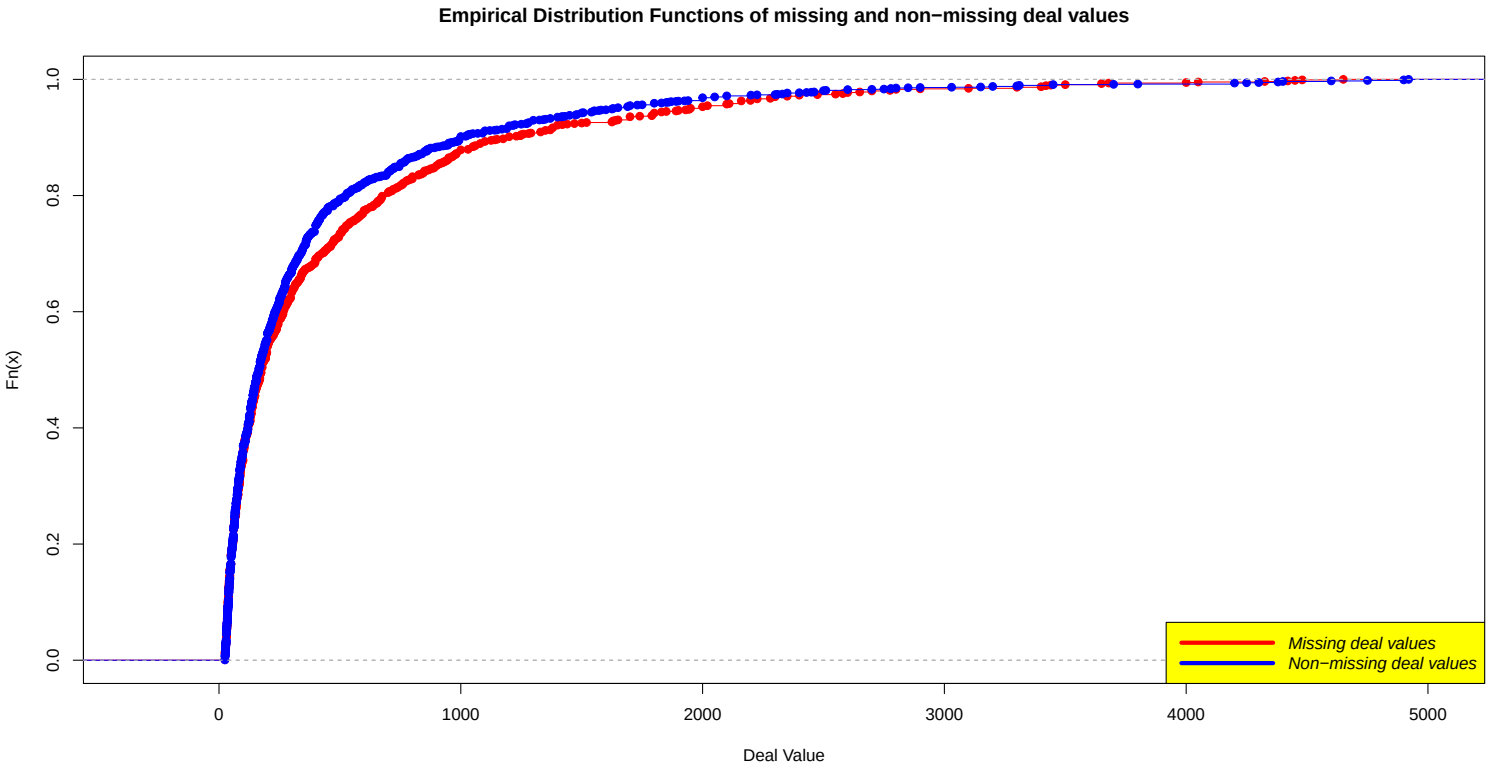
Put them together to compare



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

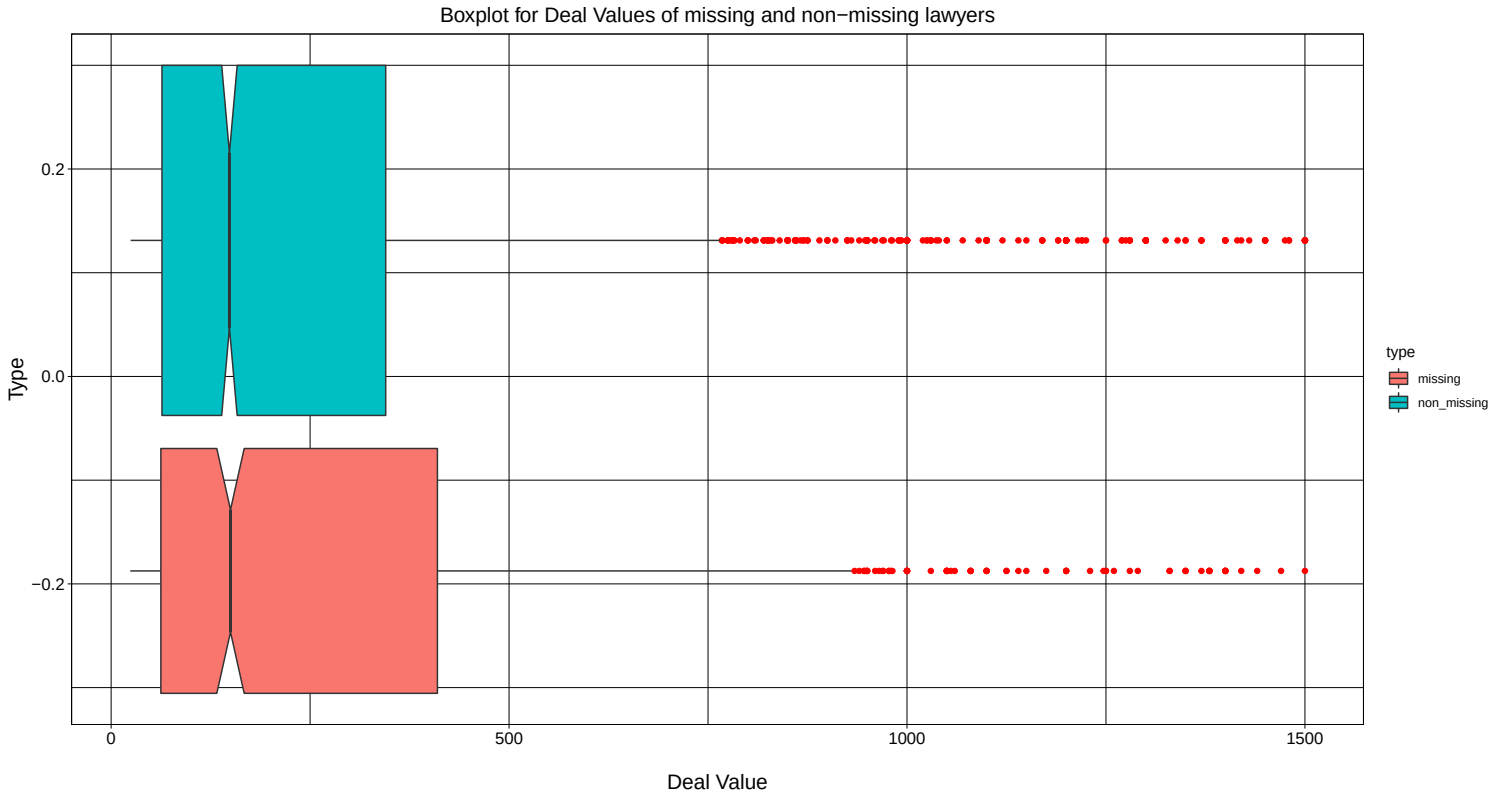
```
> miss_non_miss %>%
+     group_by(type) %>%
+     summarise(`:=`(min, min(value1, na.rm = T)), `:=`(mean, mean(value1, na.rm = T)),
+         `:=`(median, median(value1, na.rm = T)), `:=`(max, max(value1, na.rm = T)),
+         `:=`(sd, sd(value1, na.rm = T))) %>%
+     kbl(caption = "Summary Stats 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))
```

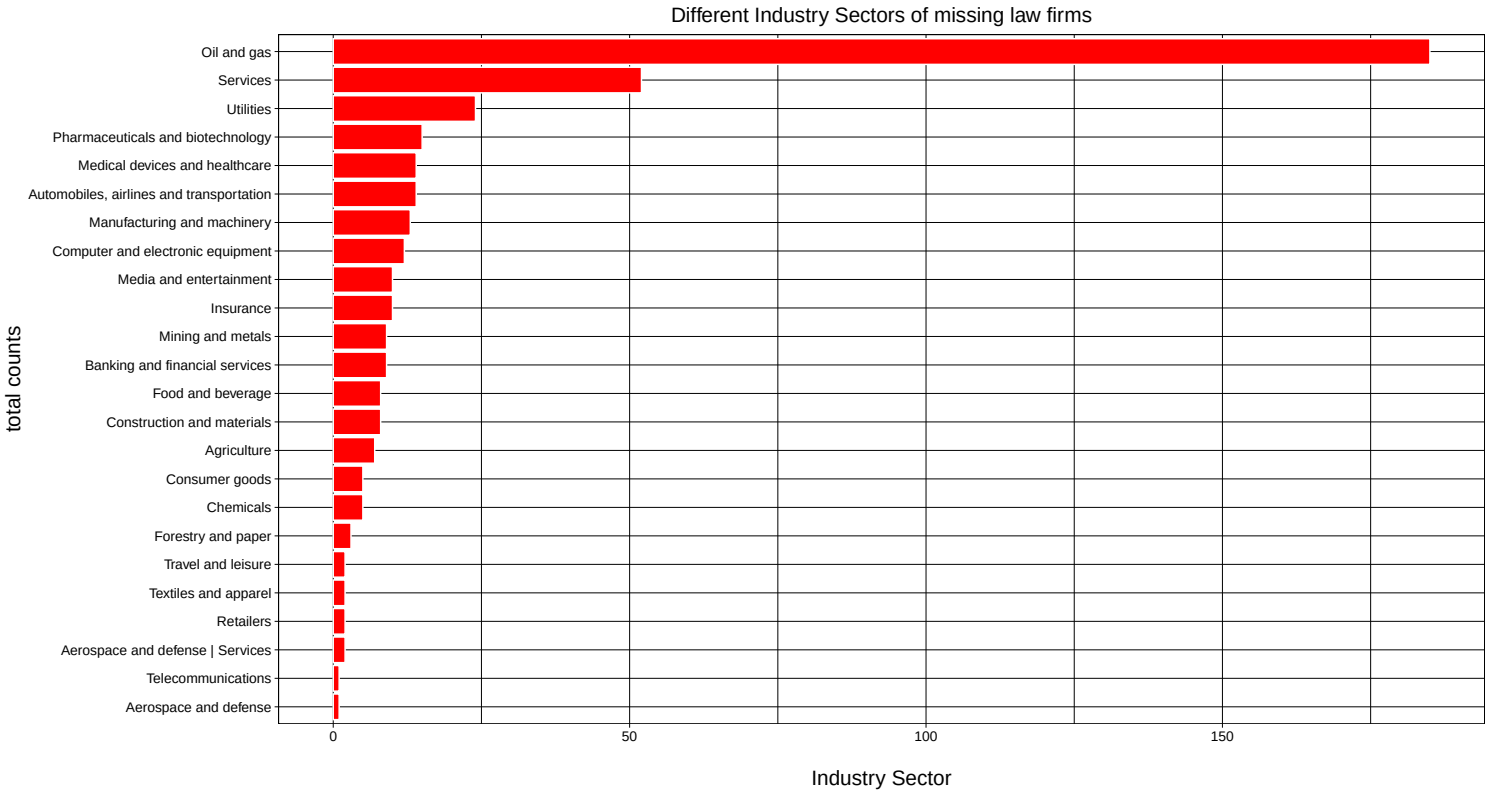


How many law firms are unknown (NA and “not disclosed”)?

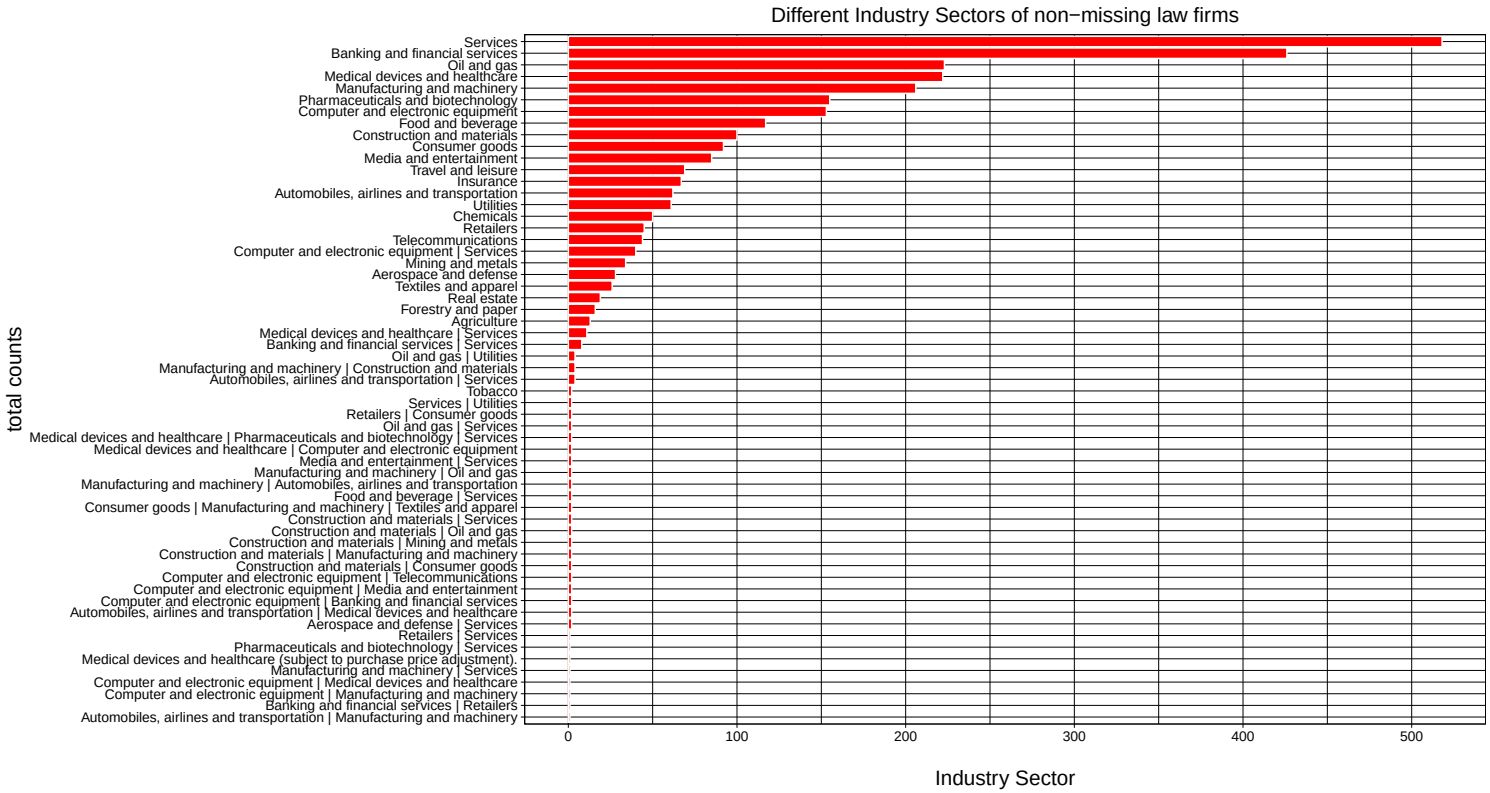
What is the distribution of the missing law firms by the size (in dollars) of the deal, year, and industry? Histogram each.

Extract corresponding dataset

Bar plots for missing law firms



Bar plots for non_missing law firms

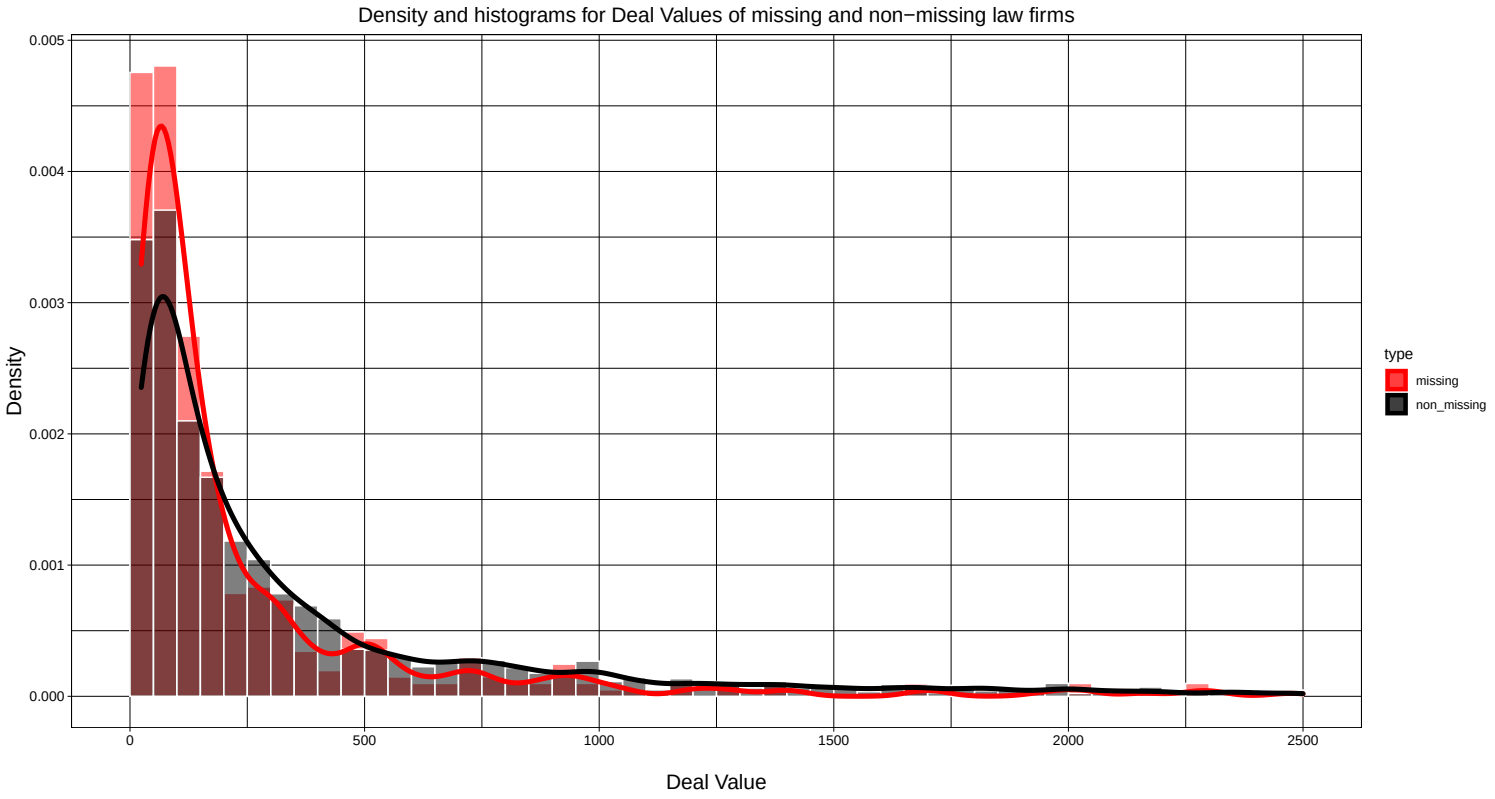


Histogram and density curves to compare them

```
> 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)
```

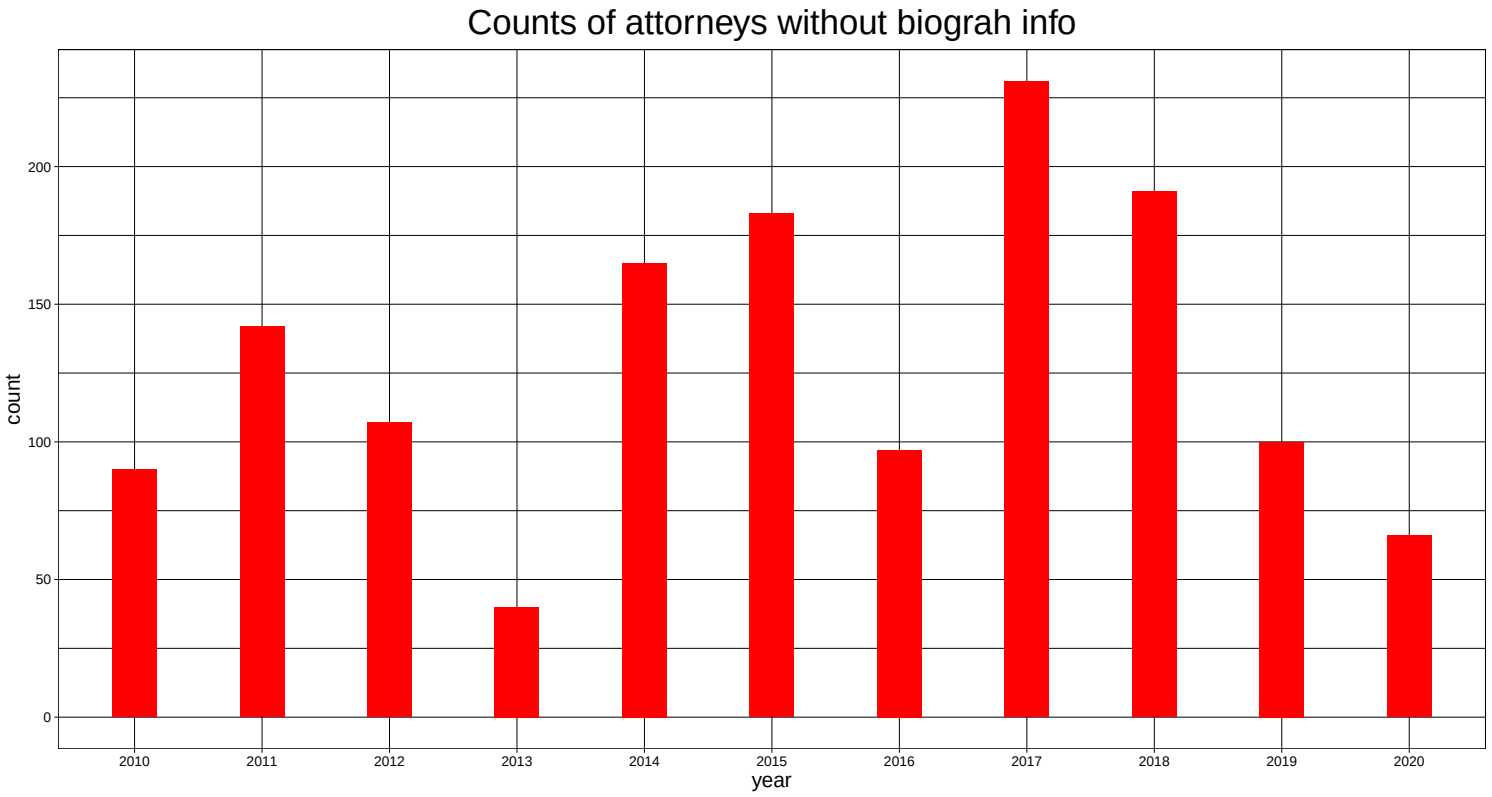
```
>
> att_miss_bio %>%
+     distinct(First, Last) %>%
+     nrow
```

```
## [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
```

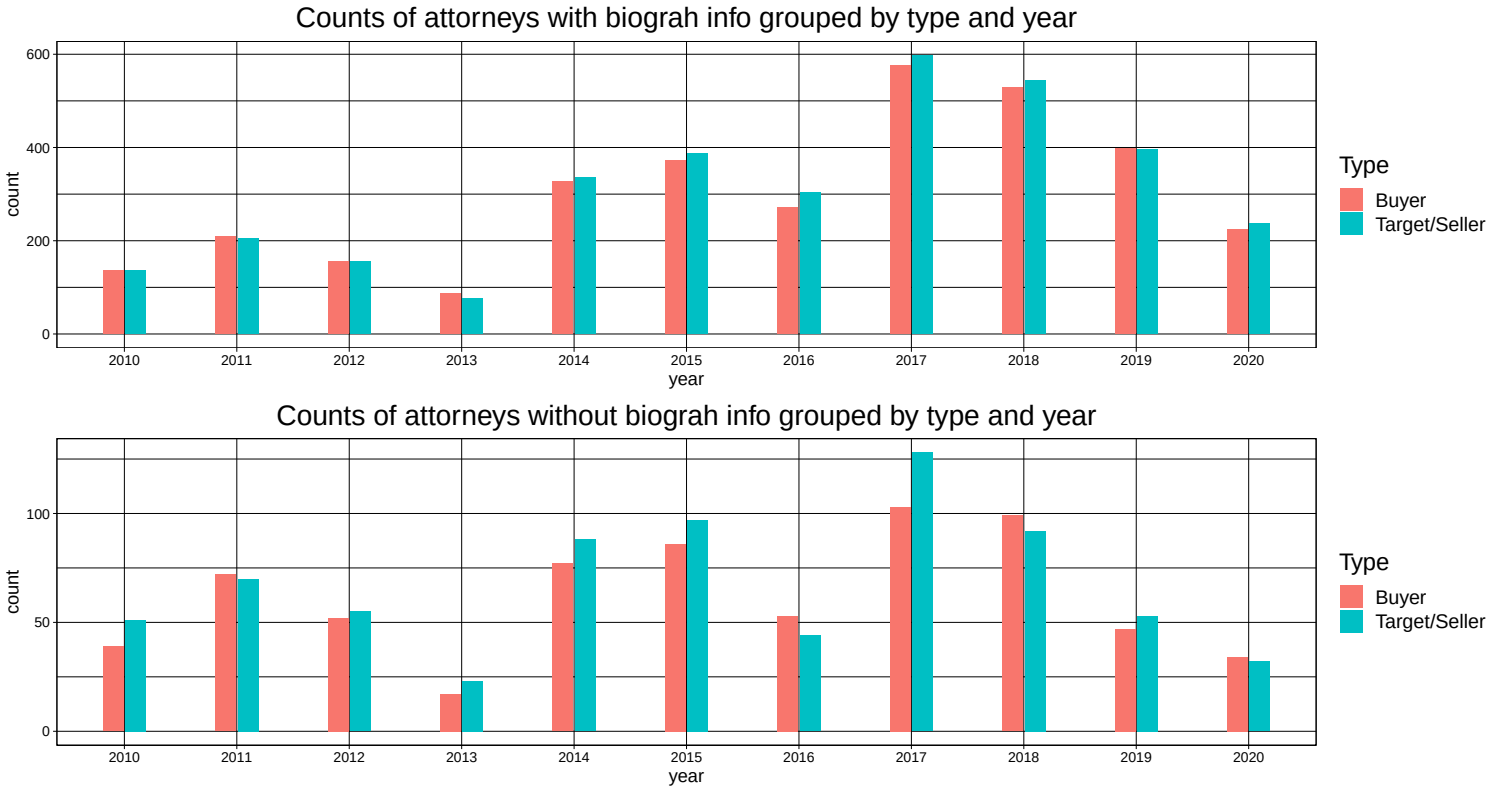
Distribution of attorneys with biograh info

```
> 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)
```

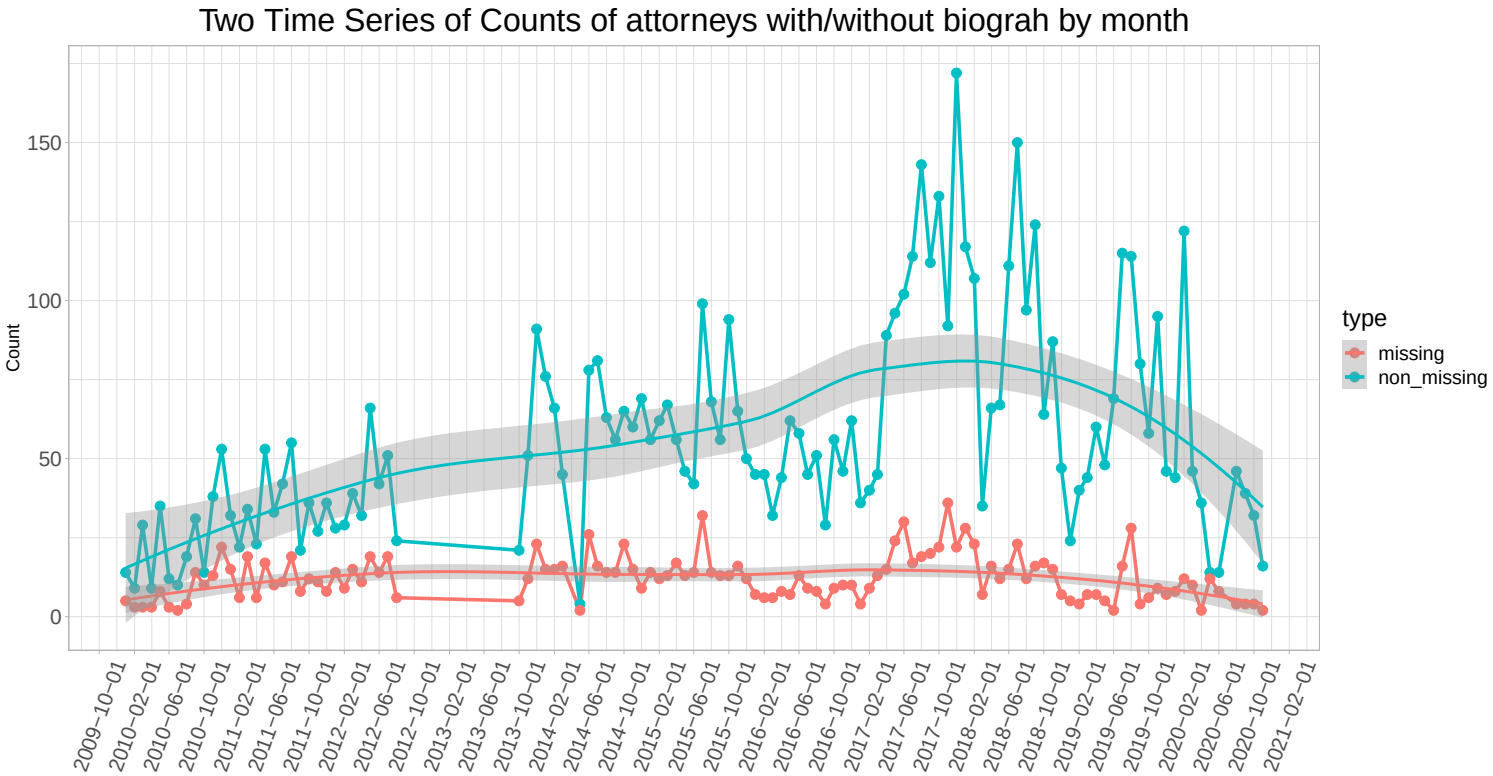
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

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. From the `loess` fitted line, we can tell that non-missing group has much fluctuation than missing group.

How many deals are missing attorneys' biographical information?

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

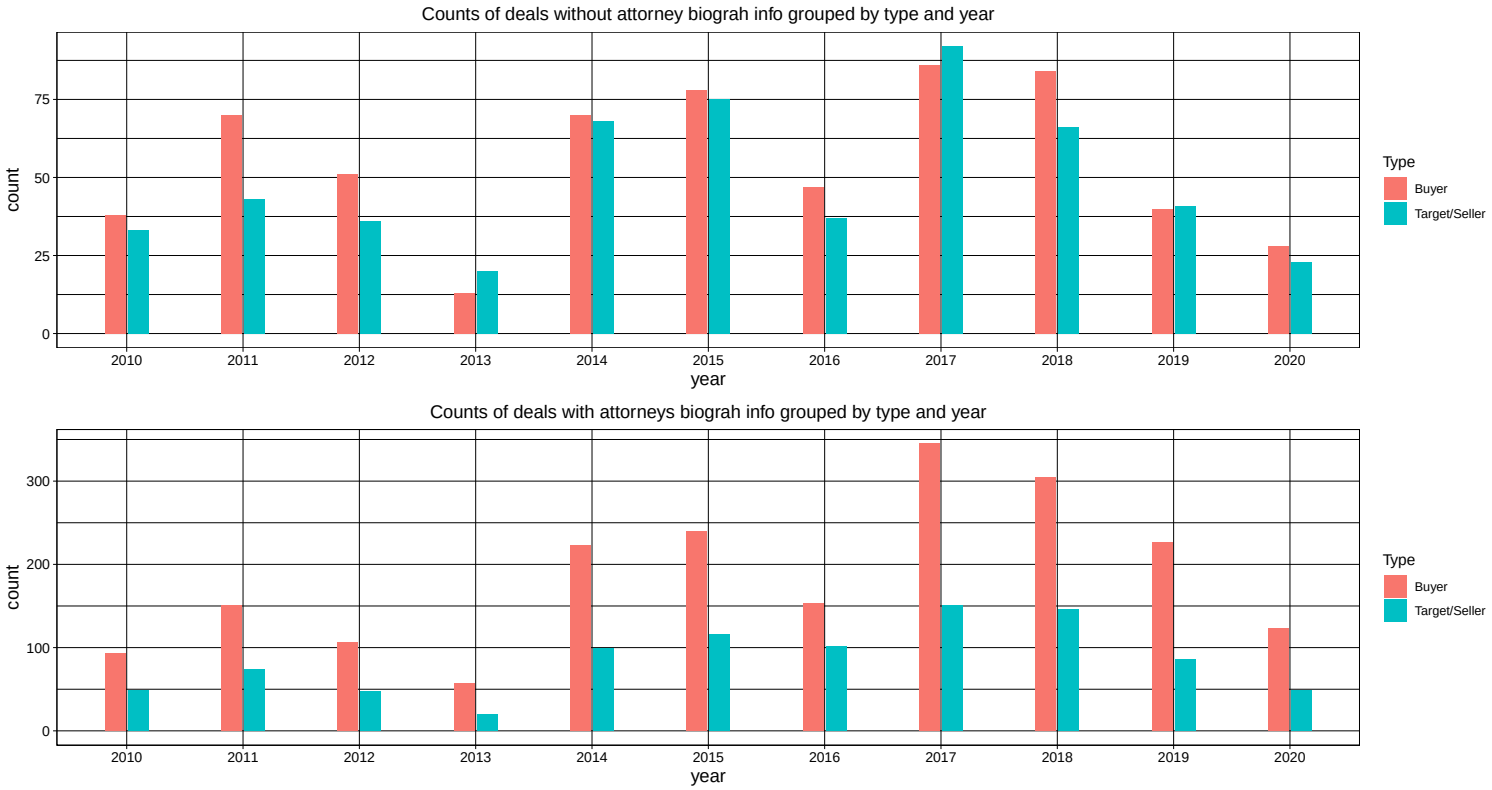
```
## [1] 1139
```

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)) %>%
+     ungroup() %>%
+     mutate(`:=`(percentage, round(n/sum(n) * 100, 2))) %>%
```

```
+     kbl(caption = "Summary Stats Table of Attorneys with Biograph", booktabs = T) %>%
+     kable_styling(latex_options = c("striped", "hold_position"))
```

Table 4: Summary Stats Table of Attorneys with Biograph

Gender	age_breaks	n	percentage
Male	45 < age <= 60	2713	50.76
Male	30 < age <= 45	1027	19.21
Male	age > 60	1011	18.91
Female	45 < age <= 60	295	5.52
Female	30 < age <= 45	213	3.99
Female	age > 60	68	1.27
Male	age <= 30	15	0.28
Female	age <= 30	3	0.06

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.

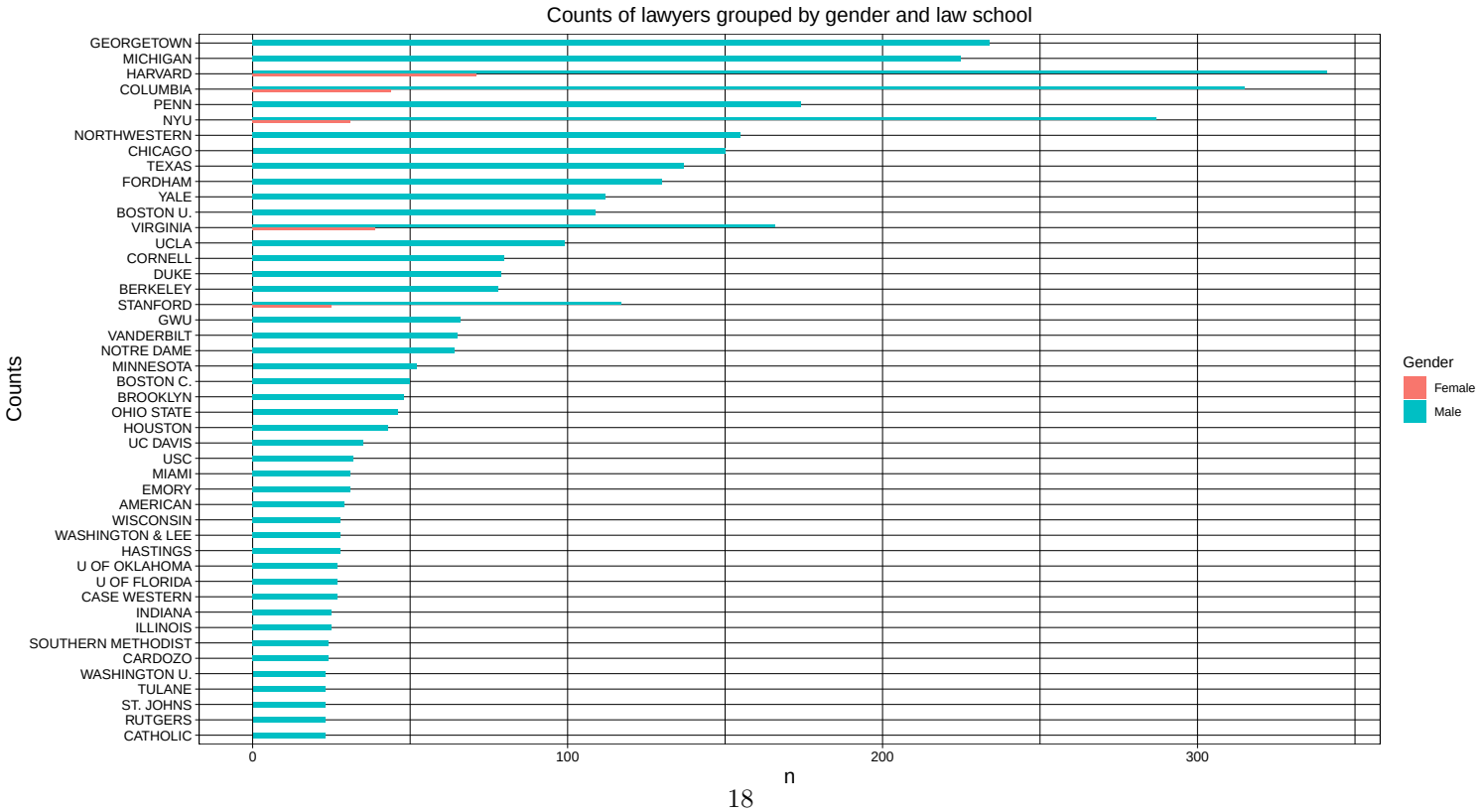
```
> dis_three_matched_stacked %>%
+     drop_na(Gender, `Law School`) %>%
+     group_by(`Law School`, Law_Firm) %>%
+     count() %>%
+     arrange(desc(n)) %>%
+     write_csv(file = "../data/deal_lawyer/distri_law_sch_firm.csv")
```

```
> 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	Percentage
MICHIGAN	Kirkland & Ellis LLP	55	1.03
NORTHWESTERN	Kirkland & Ellis LLP	53	0.99
HARVARD	Latham & Watkins LLP	42	0.79
GEORGETOWN	Skadden, Arps, Slate, Meagher & Flom LLP	39	0.73
COLUMBIA	Wachtell, Lipton, Rosen & Katz	35	0.66
YALE	Wachtell, Lipton, Rosen & Katz	34	0.64
CHICAGO	Kirkland & Ellis LLP	32	0.60
NYU	Latham & Watkins LLP	32	0.60
FORDHAM	Skadden, Arps, Slate, Meagher & Flom LLP	30	0.56
COLUMBIA	Paul, Weiss, Rifkind, Wharton & Garrison LLP	29	0.54
HARVARD	Wachtell, Lipton, Rosen & Katz	28	0.52
COLUMBIA	Latham & Watkins LLP	27	0.51
NYU	Skadden, Arps, Slate, Meagher & Flom LLP	26	0.49
TEXAS	Latham & Watkins LLP	25	0.47
HARVARD	Sidley Austin LLP	24	0.45
NYU	Wachtell, Lipton, Rosen & Katz	24	0.45
PENN	Wachtell, Lipton, Rosen & Katz	24	0.45
COLUMBIA	Sullivan & Cromwell LLP	23	0.43
NYU	Willkie Farr & Gallagher LLP	23	0.43
UCLA	Jones Day	22	0.41
COLUMBIA	Kirkland & Ellis LLP	21	0.39
U OF OKLAHOMA	Alston & Bird LLP	21	0.39
HARVARD	Kirkland & Ellis LLP	20	0.37
MICHIGAN	Latham & Watkins LLP	20	0.37
BOSTON U.	Ropes & Gray LLP	19	0.36
HARVARD	Ropes & Gray LLP	19	0.36
NYU	Kirkland & Ellis LLP	19	0.36
TEXAS	Vinson & Elkins LLP	19	0.36
BOSTON U.	Goodwin Procter LLP	18	0.34
MINNESOTA	Faegre Baker Daniels LLP	18	0.34



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`) %>%
+     mutate(`:=`(percentage, round(n/sum(n) * 100, 2))) %>%
+     filter(n >= 10) %>%
+     kbl(caption = "Industry Sector Count and Percentage", booktabs = T) %>%
+     kable_styling(latex_options = c("striped", "hold_position"))
```

Table 6: Industry Sector Count and Percentage

Industry sector	n	percentage
Services	1199	18.01
Banking and financial services	827	12.42
Medical devices and healthcare	531	7.97
Oil and gas	498	7.48
Manufacturing and machinery	455	6.83
Pharmaceuticals and biotechnology	389	5.84
Computer and electronic equipment	354	5.32
Food and beverage	282	4.23
Construction and materials	223	3.35
Consumer goods	203	3.05
Media and entertainment	181	2.72
Insurance	175	2.63
Travel and leisure	167	2.51
Utilities	135	2.03
Automobiles, airlines and transportation	121	1.82
Telecommunications	115	1.73
Retailers	108	1.62
Chemicals	99	1.49
Computer and electronic equipment | Services	99	1.49
Mining and metals	78	1.17
Aerospace and defense	72	1.08
Textiles and apparel	60	0.90
Forestry and paper	34	0.51
Real estate	33	0.50
Agriculture	31	0.47
Banking and financial services | Services	24	0.36
Medical devices and healthcare | Services	21	0.32
Manufacturing and machinery | Construction and materials	10	0.15
Services | Utilities	10	0.15

```
> 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() %>%
+     mutate(`:=`(percentage, round(n/sum(n) * 100, 2))) %>%
+     kbl(caption = "Counts of deals grouped by types and value\\_breaks", booktabs = T) %>%
+     kable_styling(latex_options = c("striped", "hold_position"))
```

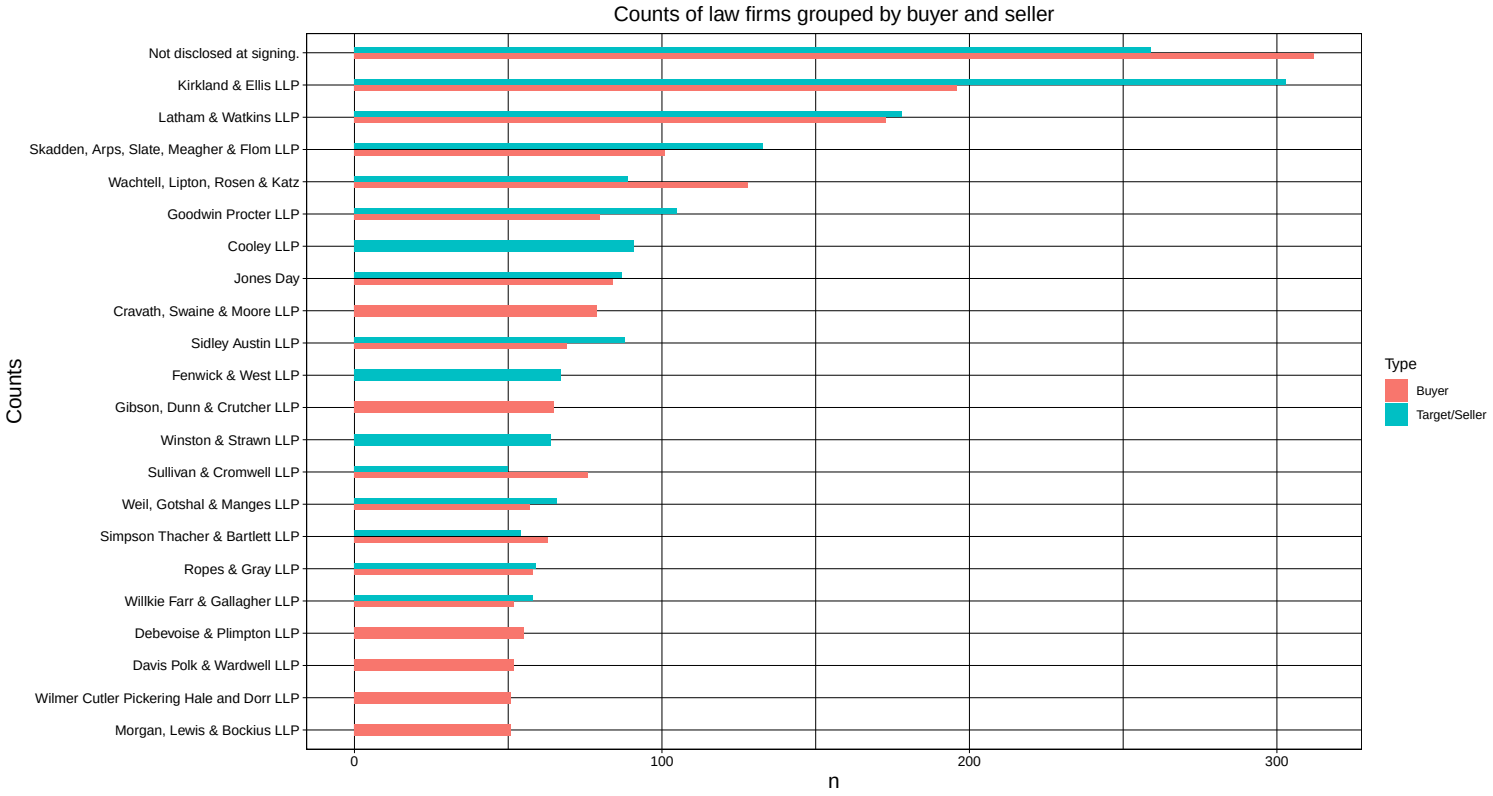
Table 7: Counts of deals grouped by types and value_breaks

Type	value_breaks	n	percentage
Target/Seller	value <= 250	1258	18.90
Buyer	value <= 250	1247	18.74
Target/Seller	250 < value <= 500	1140	17.13
Buyer	250 < value <= 500	1128	16.95
Target/Seller	1500 < value	975	14.65
Buyer	1500 < value	907	13.63

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 on Gender

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.

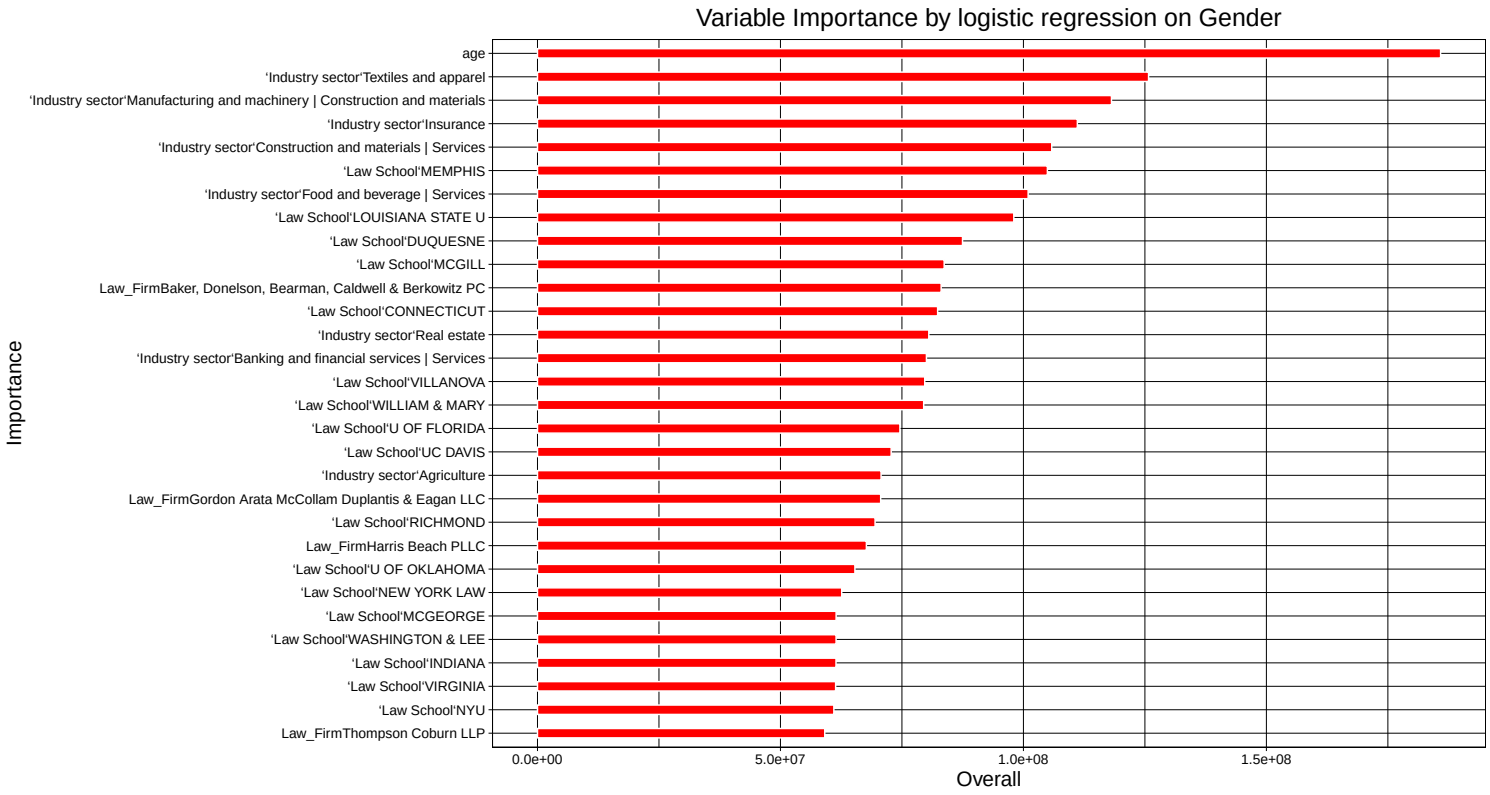
First we used Multiple Logistic Regression with all variables

```
> logit <- glm(Gender ~ ., data = dat, family = "binomial")
```

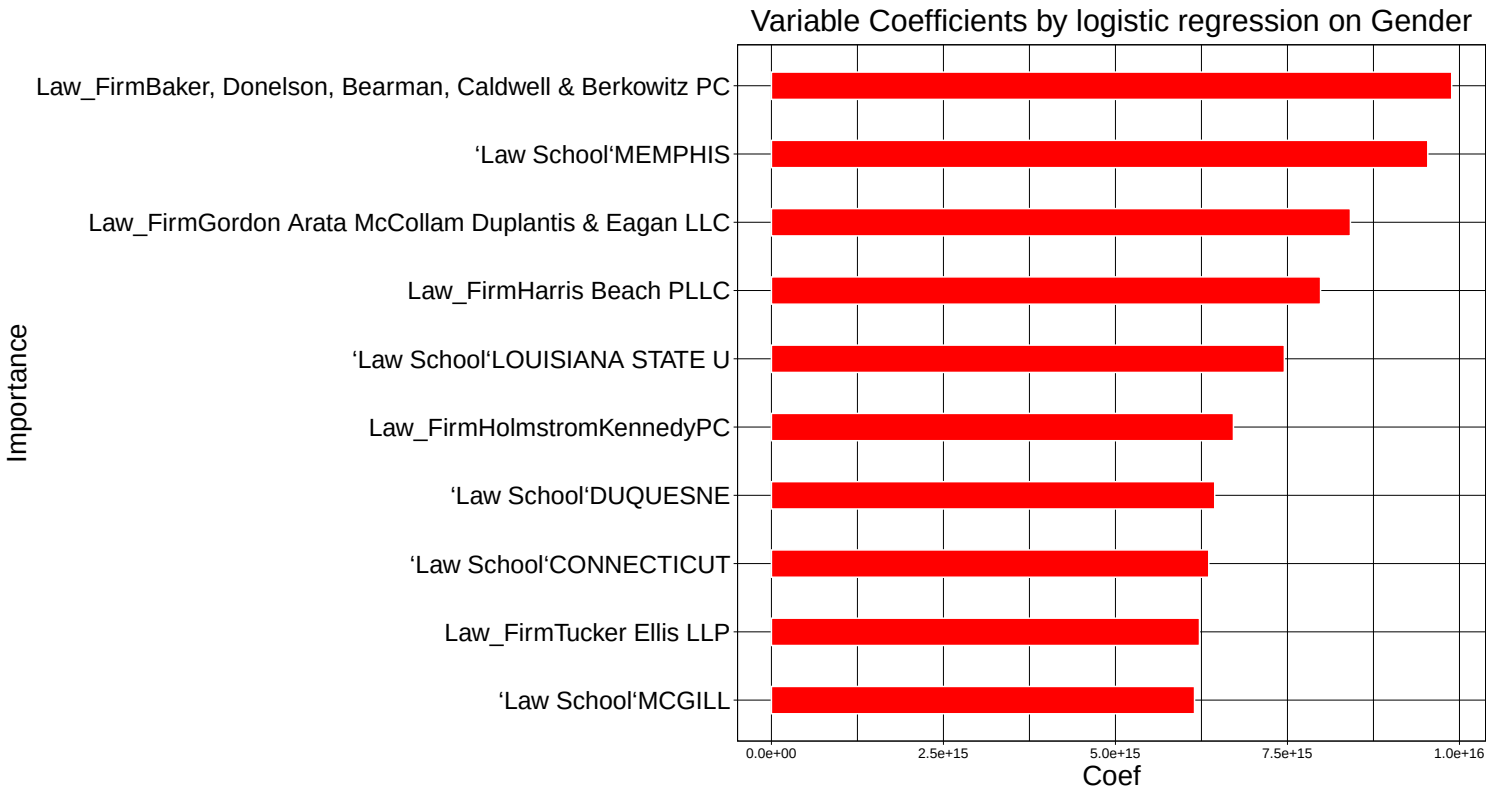
```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Selecting by Overall
```




```
## Selecting by Coef
```



age coefficients

```
> data.frame(Coef = logit$coefficients) %>%
+     rownames_to_column(var = "variable") %>%
+     filter(variable == "age")
```

```
##   variable          Coef
## 1      age -3.411114e+13
```

Law_school coefficient signs

```
> data.frame(Coef = logit$coefficients) %>%
+     rownames_to_column(var = "variable") %>%
+     filter(grepl("Law School", variable)) %>%
+     filter(Coef > 0) %>%
+     nrow
```

```
## [1] 128
```

```
> data.frame(Coef = logit$coefficients) %>%
+     rownames_to_column(var = "variable") %>%
+     filter(grepl("Law School", variable)) %>%
+     filter(Coef < 0) %>%
+     nrow
```

```
## [1] 18
```

Law_Firm coefficient signs

```
> data.frame(Coef = logit$coefficients) %>%
+     rownames_to_column(var = "variable") %>%
+     filter(grepl("Law_Firm", variable)) %>%
+     filter(Coef > 0) %>%
+     nrow
```

```
## [1] 174
```

```
> data.frame(Coef = logit$coefficients) %>%
+     rownames_to_column(var = "variable") %>%
+     filter(grepl("Law_Firm", variable)) %>%
+     filter(Coef < 0) %>%
+     nrow
```

```
## [1] 296
```

Industry Sector coefficient signs

```
> data.frame(Coef = logit$coefficients) %>%
+     rownames_to_column(var = "variable") %>%
+     filter(grepl("Industry", variable)) %>%
+     filter(Coef > 0) %>%
+     nrow
```

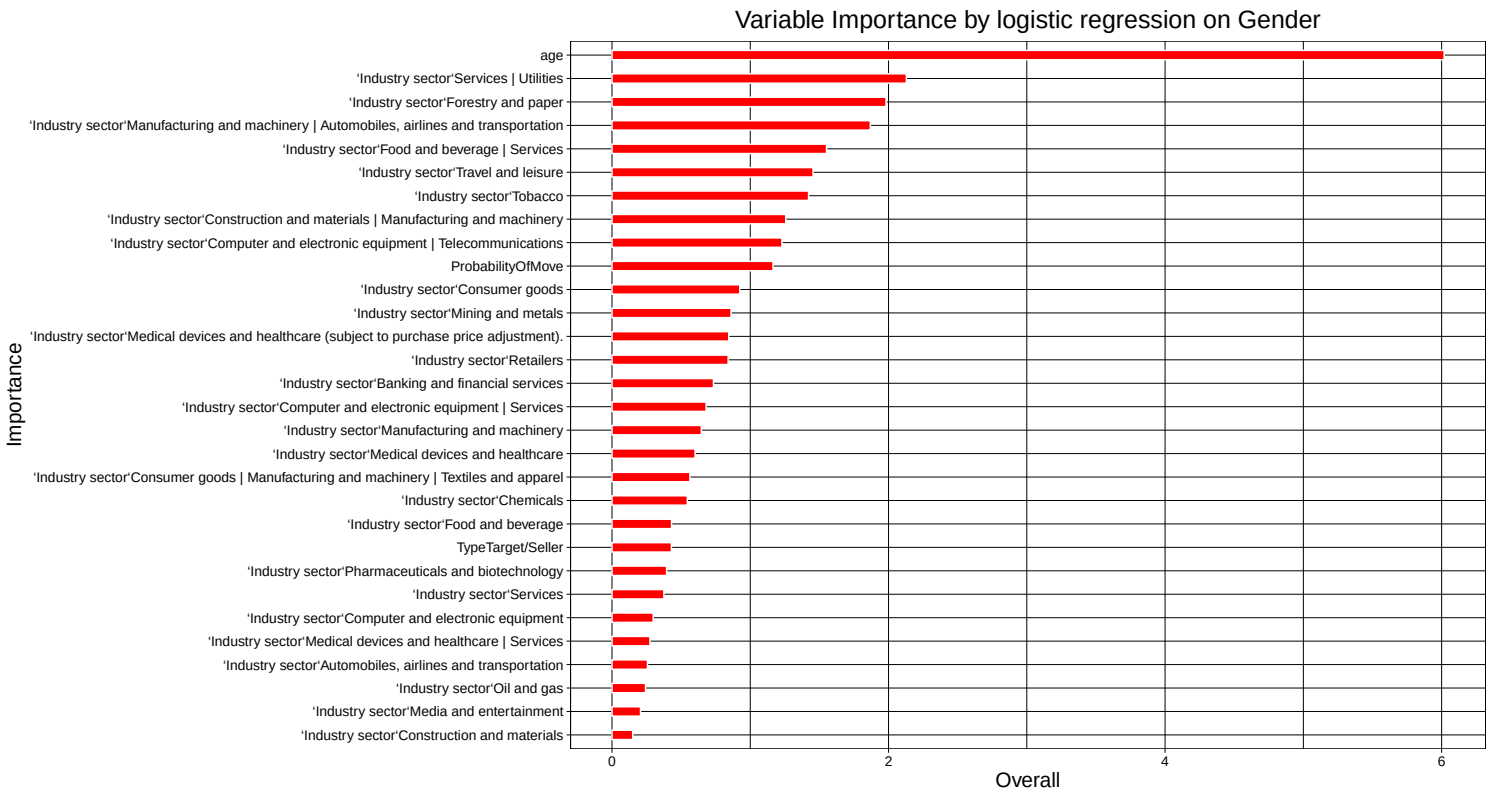
```
## [1] 17
```

```
> data.frame(Coef = logit$coefficients) %>%
+     rownames_to_column(var = "variable") %>%
+     filter(grepl("Industry", variable)) %>%
+     filter(Coef < 0) %>%
+     nrow
```

```
## [1] 30
```

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

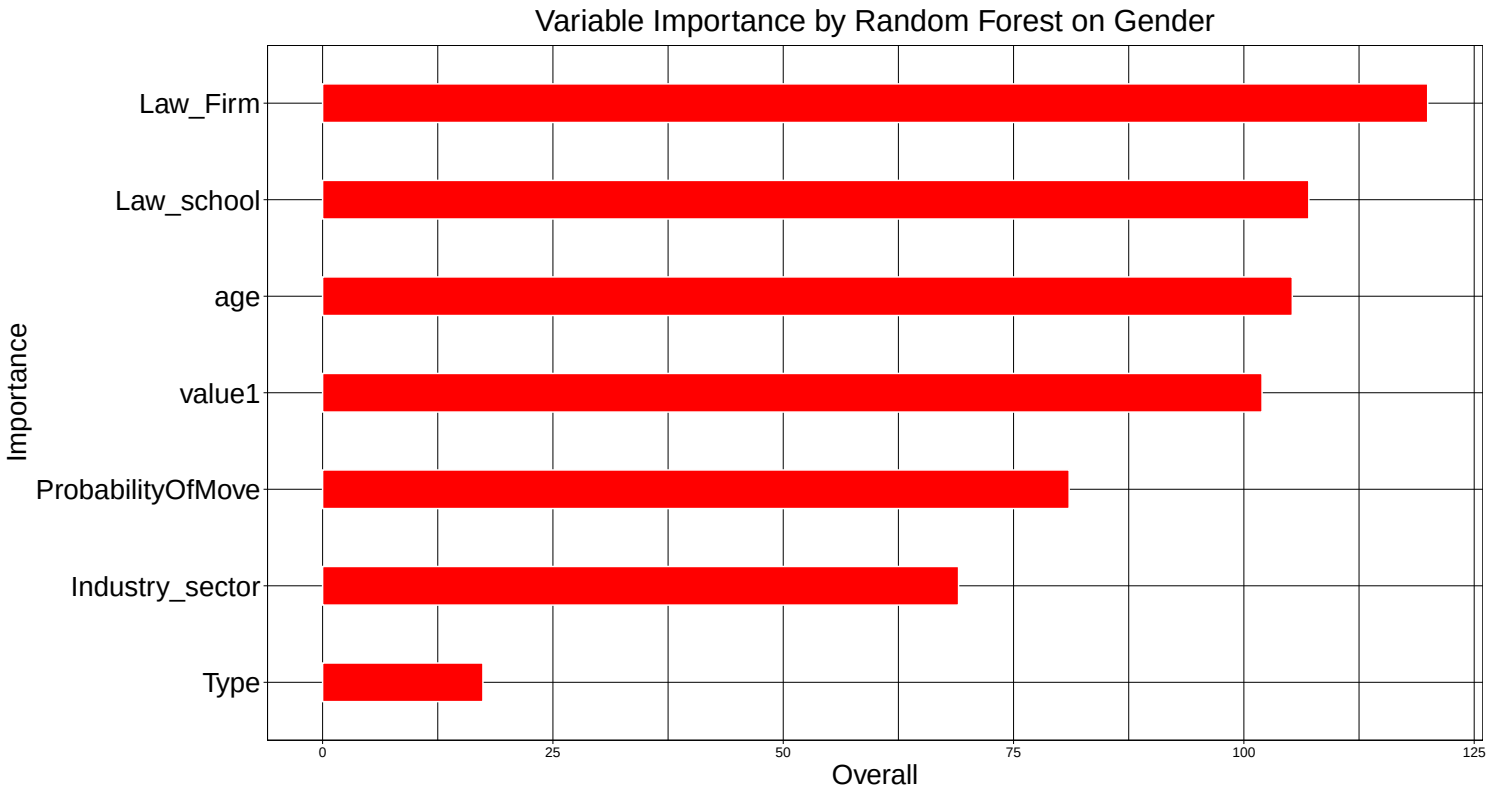
```
## Selecting by Overall
```



Random Forest with the whole variables used

```
> dat <- dat %>%
+     rename(Law_school = `Law School`, Industry_sector = `Industry sector`)
> rand <- randomForest(Gender ~ ., data = dat, na.action = na.omit)
```

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.

Predict Gender for attorneys without biograh info

```
> name_gender_dataset <- read_csv("../data/gender/name_gender_dataset.csv")
> # View(name_gender_dataset)
```

We collect baby names from 1940 to 1990 provided by SSA since this year range covers most lawyer birth years.

We displayed top 10 rows.

```
> head(name_1940_1990, 10) %>%
+     kbl(caption = "Baby Names from 1940 to 1990 provided by SSA", booktabs = T) %>%
+     kable_styling(latex_options = c("striped", "hold_position"))
```

```
> gender_UCI <- name_gender_dataset %>%
+     group_by(Name) %>%
```

Table 9: Baby Names from 1940 to 1990 provided by SSA

name	sex1	prop1
Cayce	unclear	0.25
Shawndale	unclear	0.25
Abiola	unclear	0.25
Dayan	unclear	0.25
Shondale	unclear	0.25
Lexus	unclear	0.25
Kippie	unclear	0.25
Kona	unclear	0.25
Terrylee	unclear	0.25
Damie	unclear	0.25

```
+     mutate(`:=`(prop, Count/sum(Count))) %>%
+     filter(prop >= 0.25) %>%
+     group_by(Name) %>%
+     mutate(`:=`(sex, case_when(n() == 1 ~ Gender, n() == 2 ~ "unclear"))) %>%
+     distinct(Name, sex, .keep_all = T) %>%
+     select(Name, sex, prop)
```

We used `Gender by Name Data Set` from https://archive.ics.uci.edu/ml/datasets/Gender+by+Name# as the supplementary to teh first babynames data. We displayed top 10 rows.

```
> head(gender_UCI, 10) %>%
+     kbl(caption = "Names and Genders from provided by UCI", booktabs = T) %>%
+     kable_styling(latex_options = c("striped", "hold_position"))
```

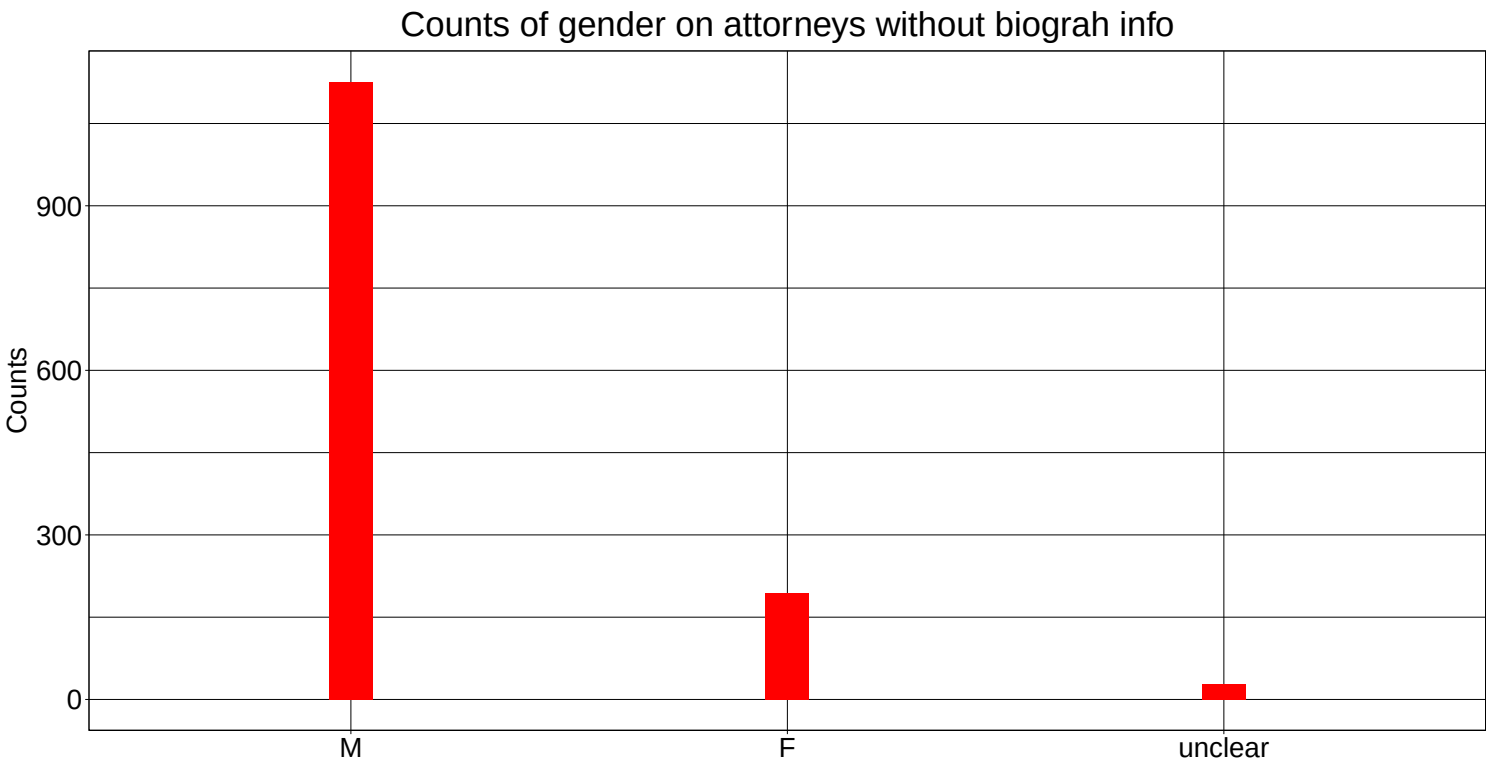
Table 10: Names and Genders from provided by UCI

Name	sex	prop
James	M	0.9955028
John	M	0.9958079
Robert	M	0.9958756
Michael	M	0.9950335
William	M	0.9961562
Mary	F	0.9963426
David	M	0.9964533
Joseph	M	0.9959029
Richard	M	0.9963096
Charles	M	0.9948446

We match attorney without biograh info using these two new gender prediction by first name dataset.

```
> att_miss_with_gender <- att_miss_bio %>%
+     left_join(name_1940_1990, by = c(First = "name")) %>%
+     left_join(gender_UCI, by = c(First = "Name")) %>%
+     mutate(`:=`(gender, case_when(is.na(sex1) ~ sex, !is.na(sex1) ~ sex1)))
```

Make some plots on Gender



From this plot we can tell that male are dominant without bigraph info.

Two-Proportions Z-Test of Gender between missing versus non-missing

The two-proportions z-test is used to compare two observed proportions. We want to know whether the proportions of female are the same in the two groups , missing biograph versus non-missing biograph?

Let's pull up gender counts in attorneys with biograph info.

```
> dis_three_matched_stacked %>%
+     drop_na(Gender) %>%
+     group_by(Gender) %>%
+     count() %>%
+     kbl(caption = "Gender Counts of attorneys with biograph", booktabs = T) %>%
+     kable_styling(latex_options = c("striped", "hold_position"))
```

Table 11: Gender Counts of attorneys with biograph

Gender	n
Female	579
Male	4766

Research questions

Whether the observed proportion of females in group missing (Pm) is equal to the observed proportion of females in group non-missing (Pnm)?

$$H_0 : P_m = P_{nm}$$
$$H_a : P_m \neq P_{nm} \text{ (different)}$$

The z-test statistics can be calculated as follow:

$$z = \frac{P_m - P_{nm}}{\sqrt{pq/n_m + pq/n_{nm}}}$$

where p and q are the overall proportions.

Form a contingency table

Table 12: Gender Counts by missing versus non missing

Gender	missing	non_missing
Female	194	579
Male	1126	4766

```
> prop.test(x = c(194, 579), n = c(194 + 1126, 579 + 4766))

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(194, 579) out of c(194 + 1126, 579 + 4766)
## X-squared = 15.044, df = 1, p-value = 0.000105
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.01733267 0.05995565
## sample estimates:
##    prop 1    prop 2
## 0.1469697 0.1083255
```

The p-value of the test is 0.000105, which is less than the significance level $\alpha = 0.05$. We can conclude that the proportion of female is significantly different in the two groups with a p-value = 0.000105. The female proportion estimate of missing group is about 0.1469 which is greater than that of non-missing group, 0.1083. We can say that there are missing female biograph than non-missing.

Thanks for [6], [7], [1], [9], [10], [3], [11], [4], [5], [8] and gender prediction dataset from UCI [2]

References

[1] W. Chang et al. *shiny: Web Application Framework for R. R package version 0.12.1*. Computer Program. 2015. URL: http://CRAN.R-project.org/package=shiny.

[2] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[3] Marek Gagolewski. *R package stringi: Character string processing facilities*. 2020. URL: http://www.gagolewski.com/software/stringi/.

[4] Max Kuhn. *caret: Classification and Regression Training*. R package version 6.0-86. 2020. URL: https://CRAN.R-project.org/package=caret.

[5] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22. URL: https://CRAN.R-project.org/doc/Rnews/.

[6] R Core Team. "R: A Language and Environment for Statistical Computing". In: (2015). URL: http://www.R-project.org.

[7] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA, 2020. URL: http://www.rstudio.com/.

[8] Hadley Wickham. *babynames: US Baby Names 1880-2017*. R package version 1.0.1. 2021. URL: https://CRAN.R-project.org/package=babynames.

[9] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[10] Hadley Wickham et al. "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43 (2019), p. 1686. DOI: 10.21105/joss.01686.

[11] Hao Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.3.4. 2021. URL: https://CRAN.R-project.org/package=kableExtra.