Compare missing versus non-missing Statistics

Li Yuan

7/14/2021

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)
```

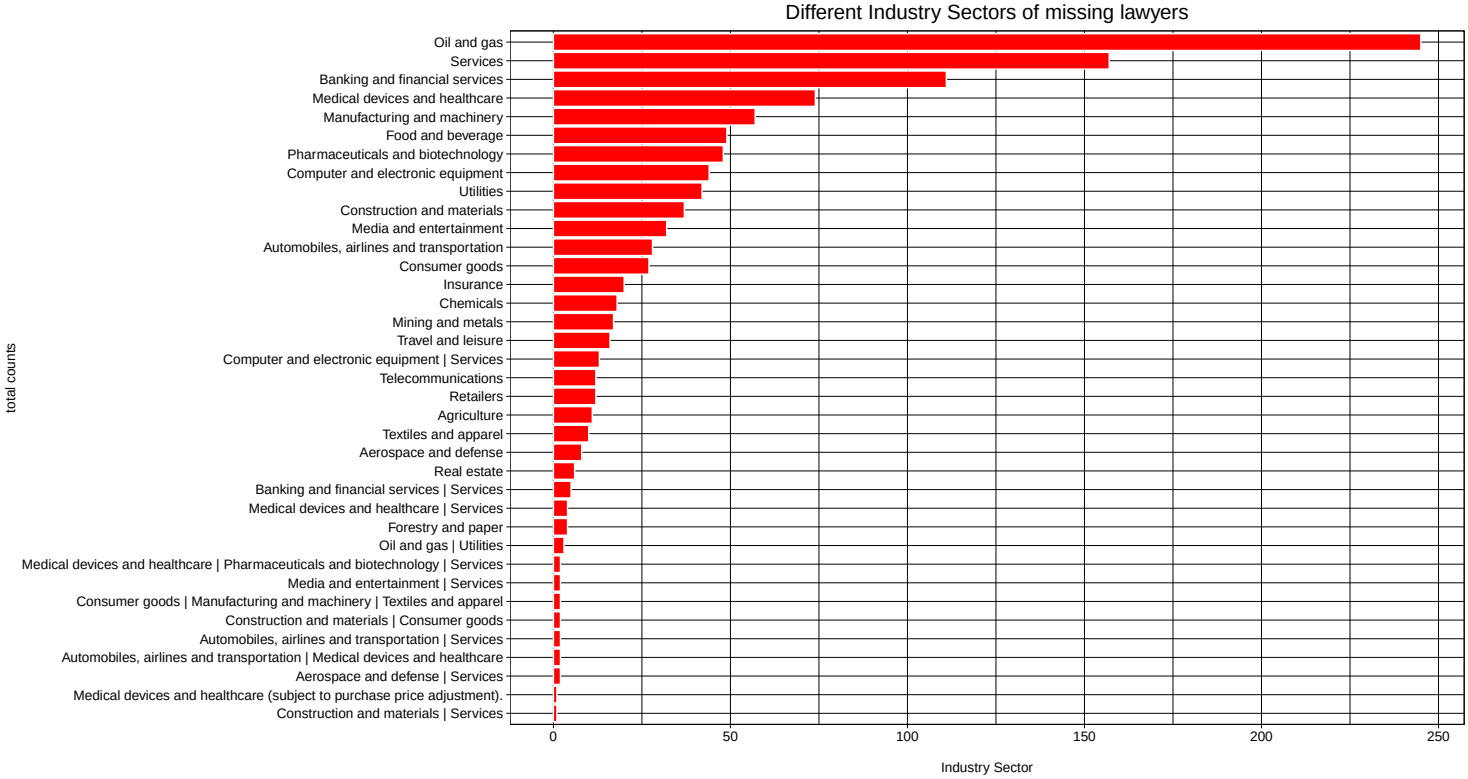
Load data

```
> 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)
```

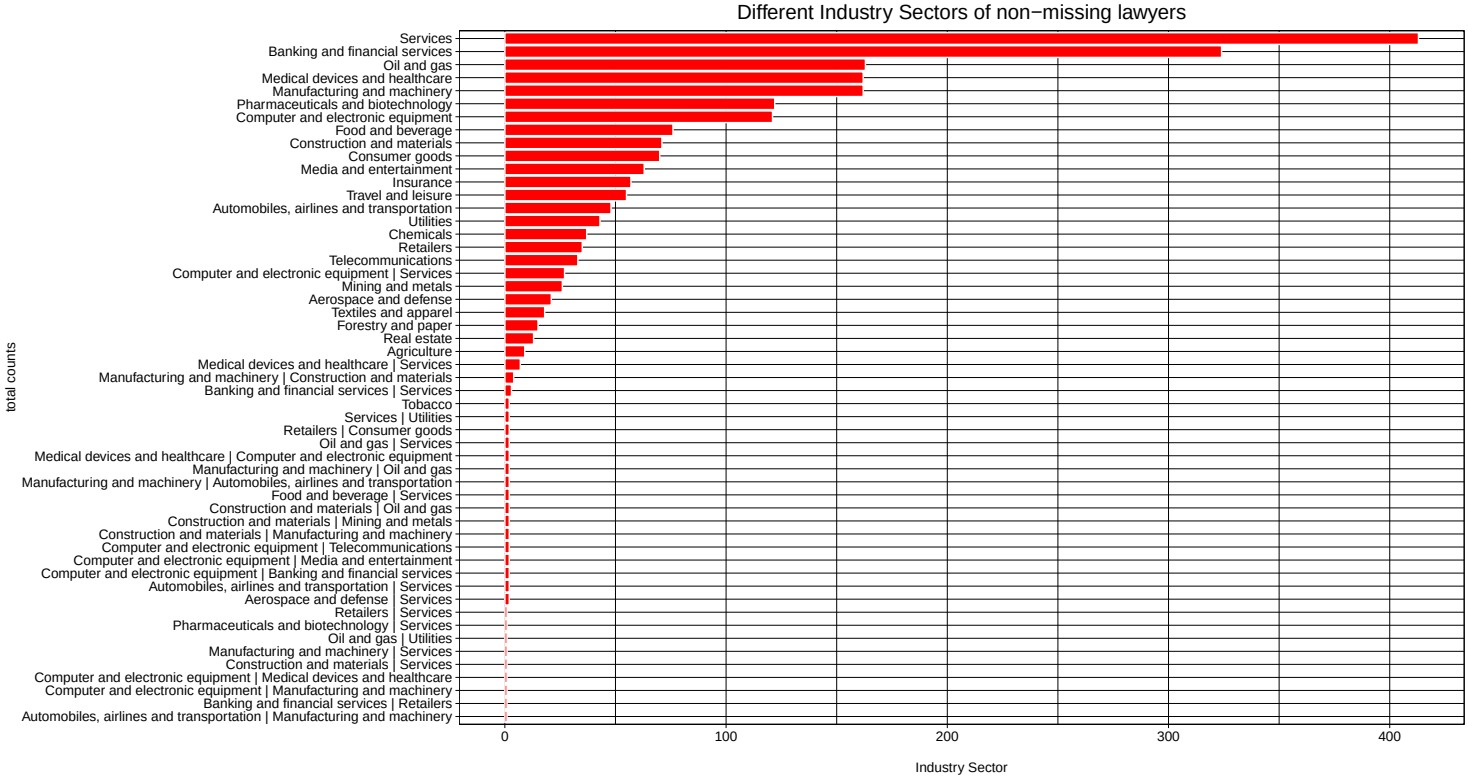
How many lawyers are unknown (NA and “not disclosed”)?

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

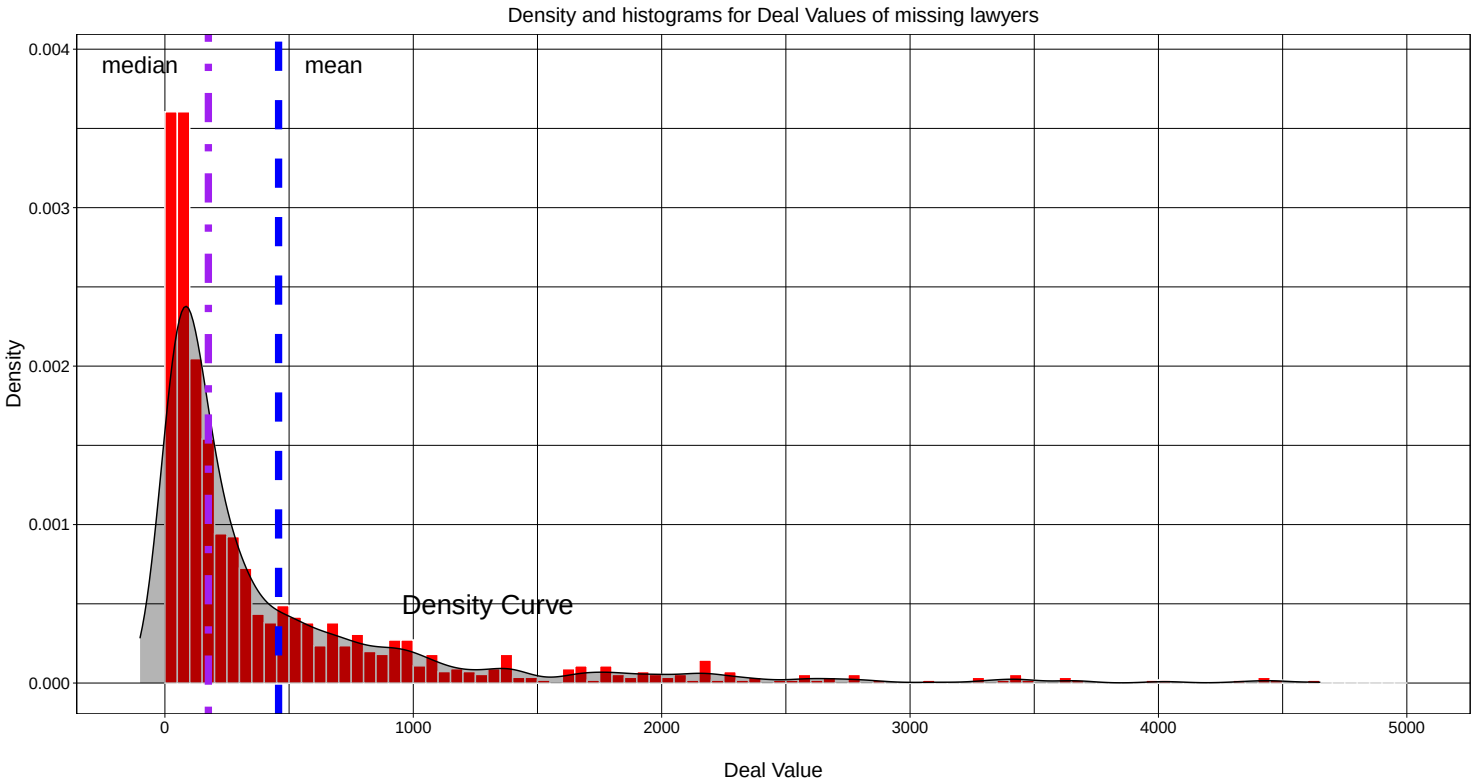
Counts of Missing Lawyers

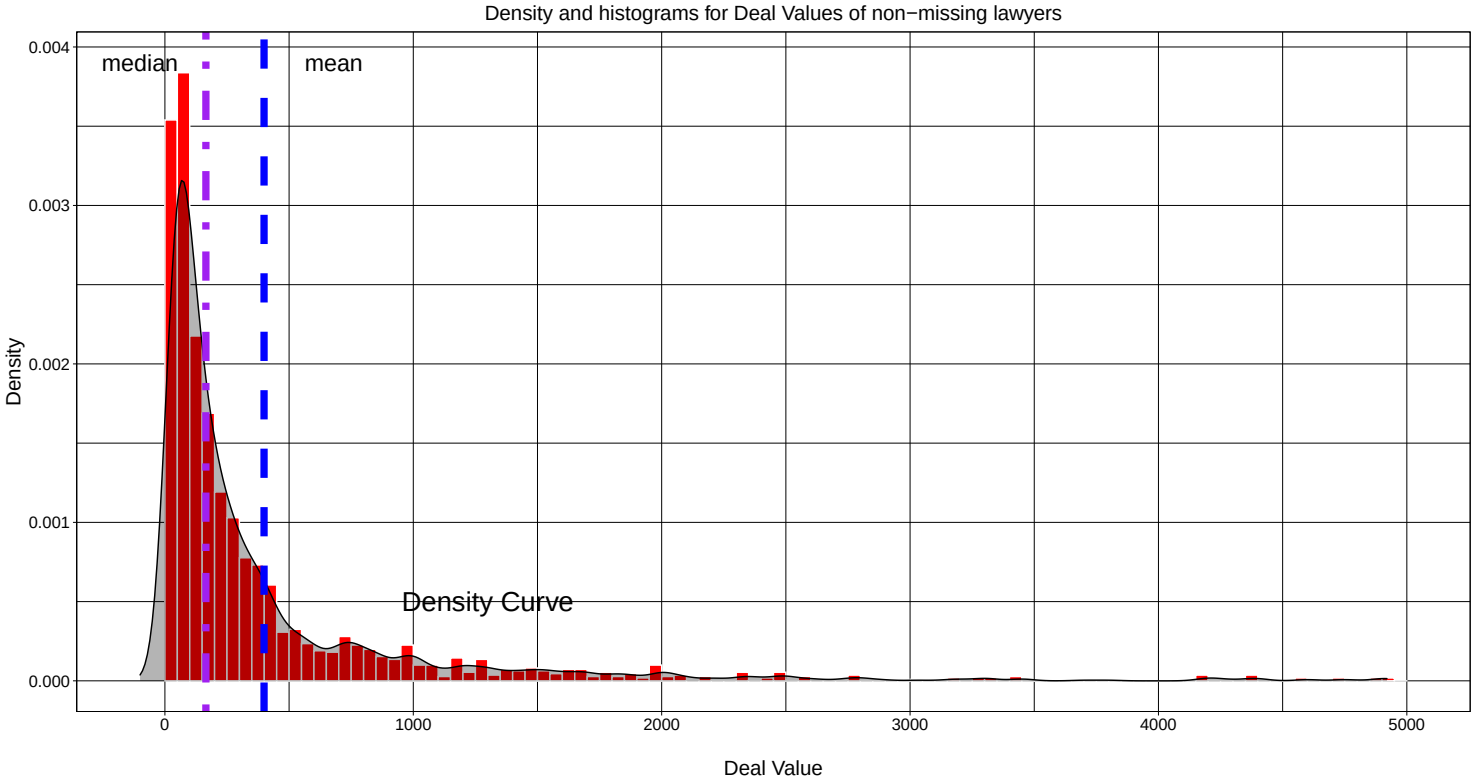


Counts of non-missing lawyers

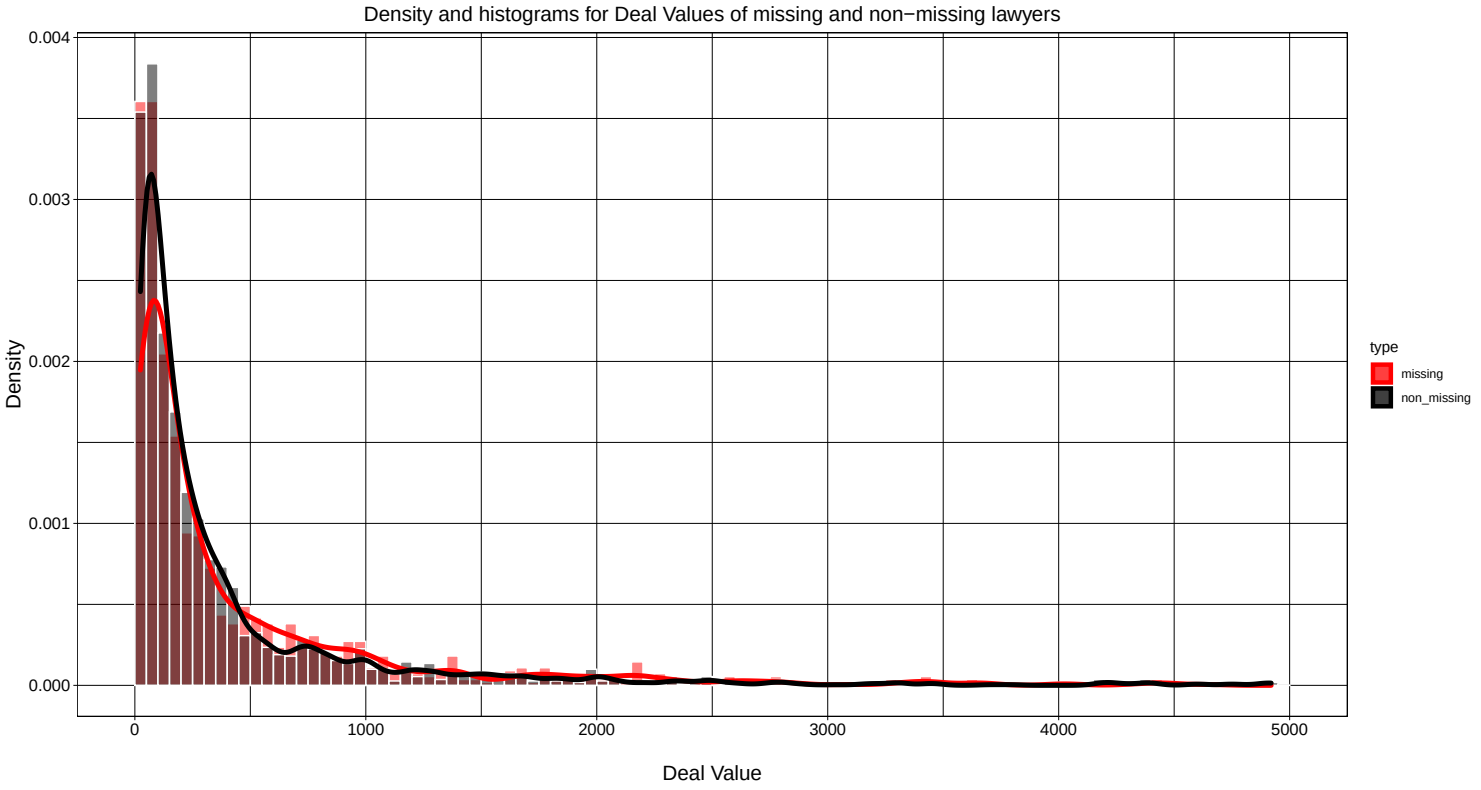


Distribution and Density curve of deal values of missing lawyers





Put them together to compare



Mean, Median and Standard Deviation of deal values

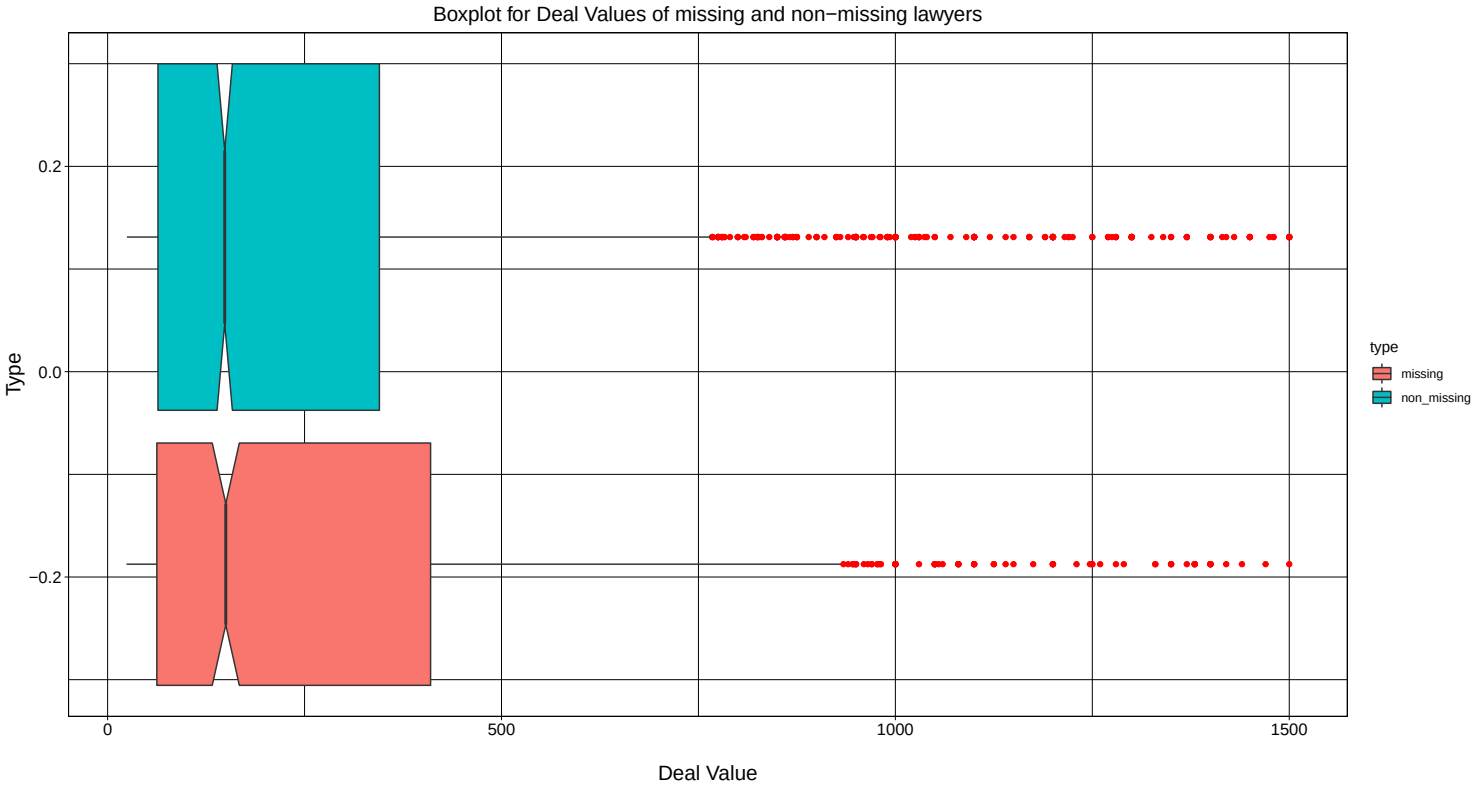
```
> miss_non_miss %>%
+     group_by(type) %>%
+     summarise(`:=`(min, min(value1, na.rm = T)), `:=`(mean, mean(value1, na.rm = T)),
+         `:=`(median, median(value1, na.rm = T)), `:=`(max, max(value1, na.rm = T)),
+         `:=`(sd, sd(value1, na.rm = T))) %>%
+     kbl(caption = "Summary Stats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

Table 1: 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))
```

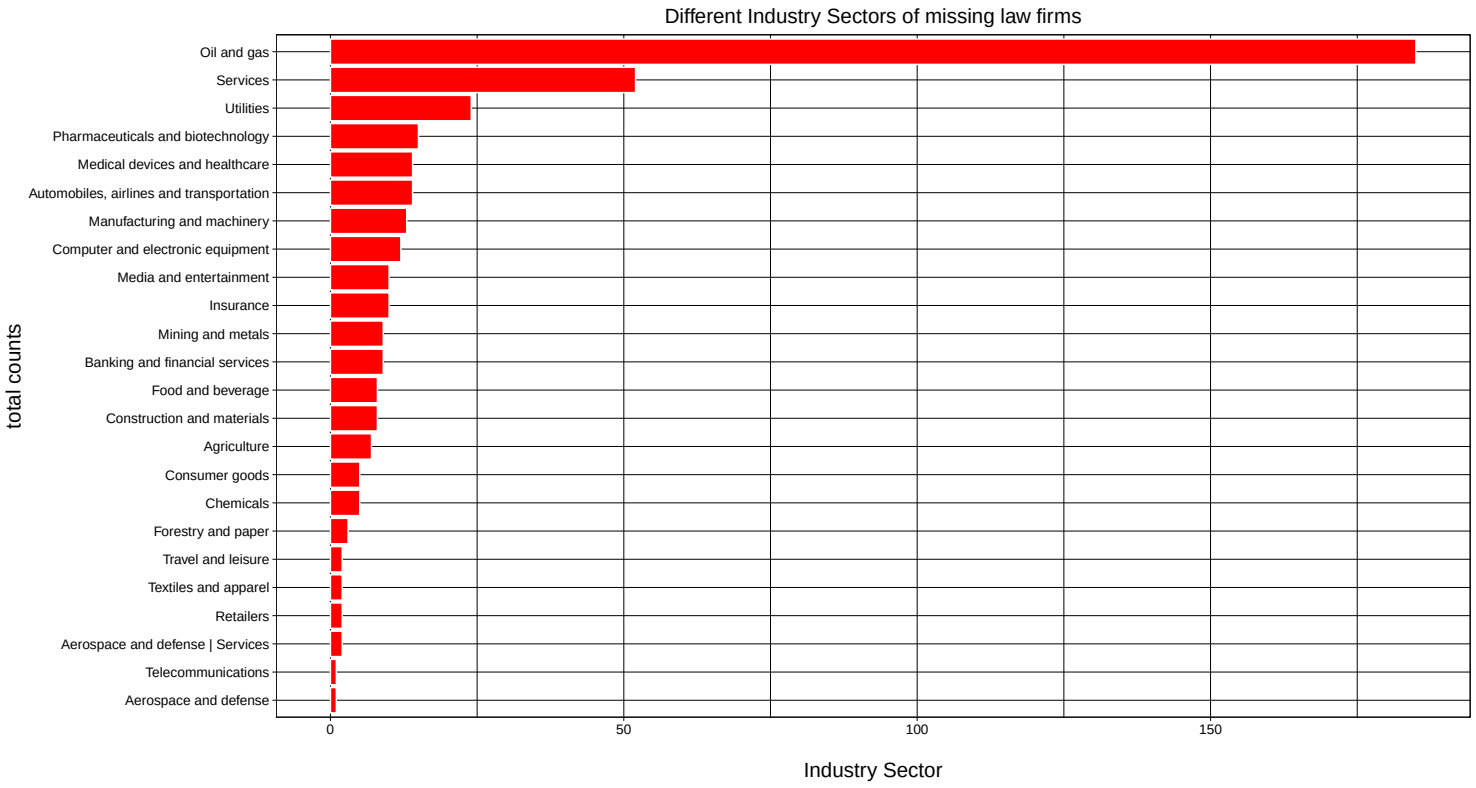


How many law firms are unknown (NA and “not disclosed”)?

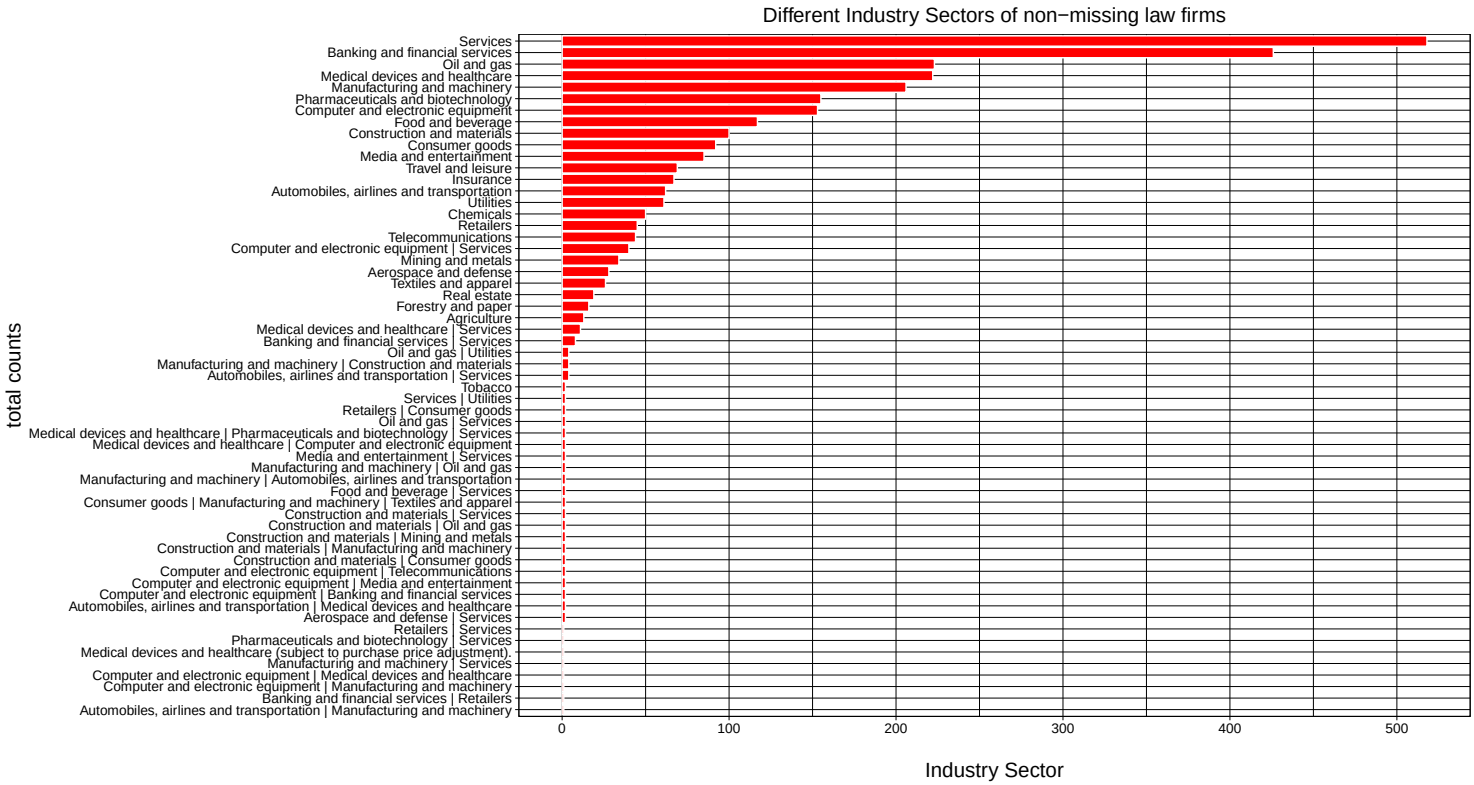
What is the distribution of the missing law firms by the size (in dollars) of the deal, year, and industry? Histogram each.

Extract corresponding dataset

Bar plots for missing law firms



Bar plots for non_missing law firms

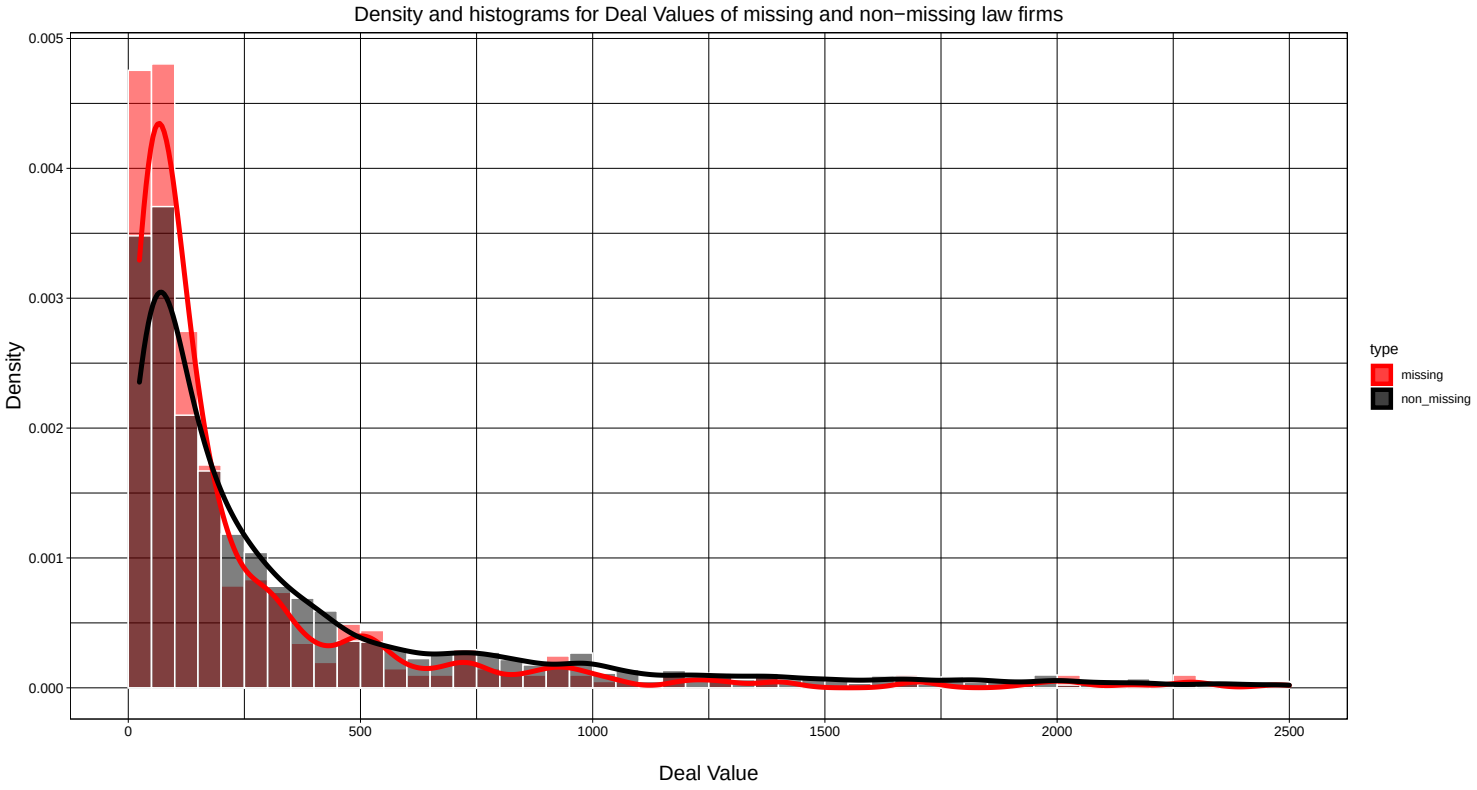


Histogram and density curves to compare them

```
> 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 2: 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.c
> 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)
>
```

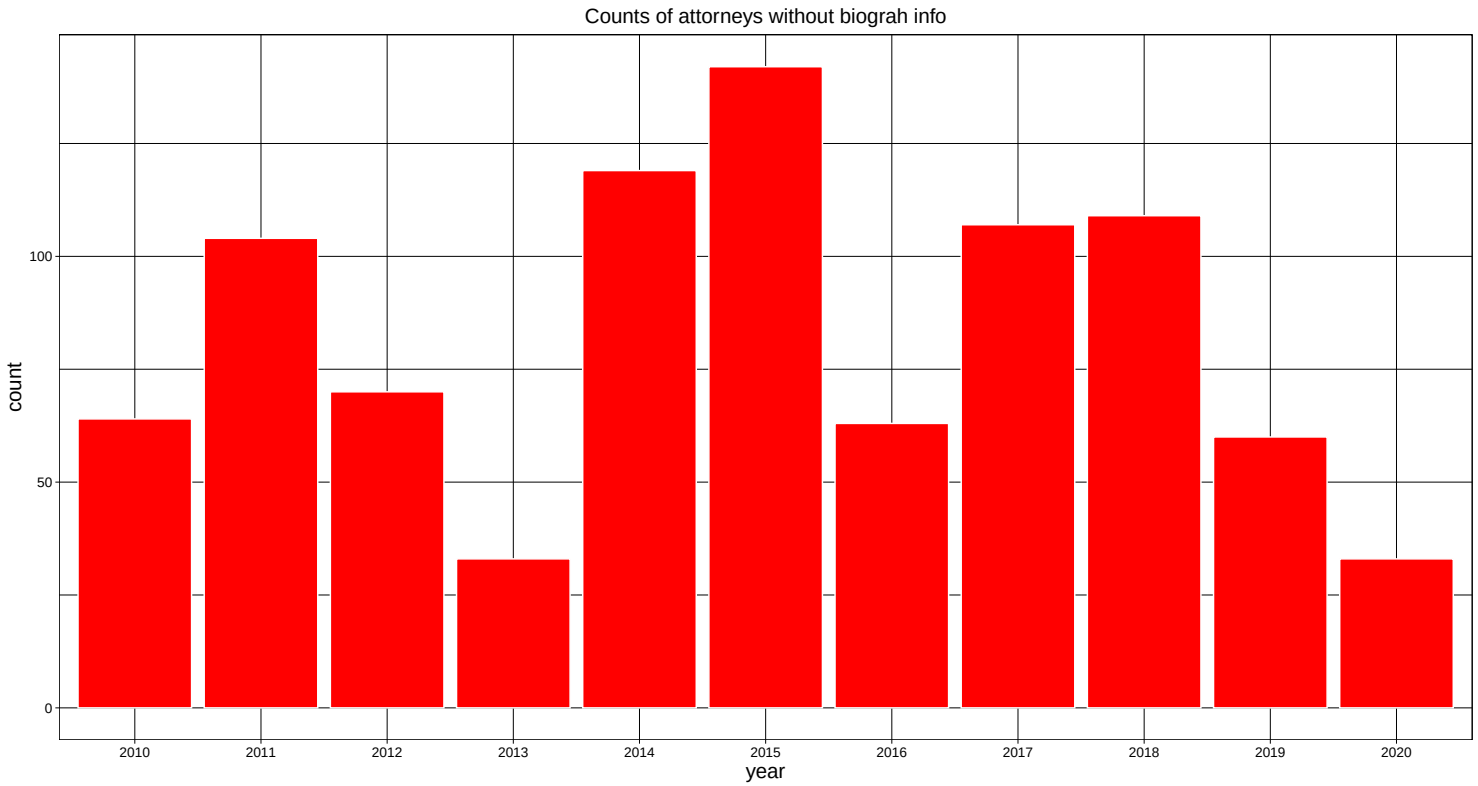
```
> att_miss_bio %>%
+     distinct(First, Last) %>%
+     nrow
```

```
## [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
```

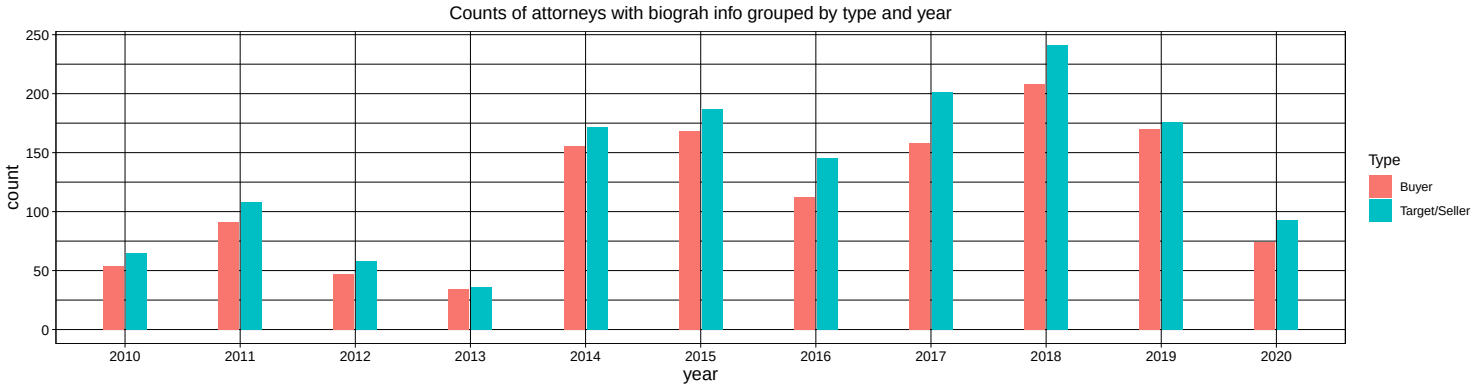
Distribution of attorneys with biograh info

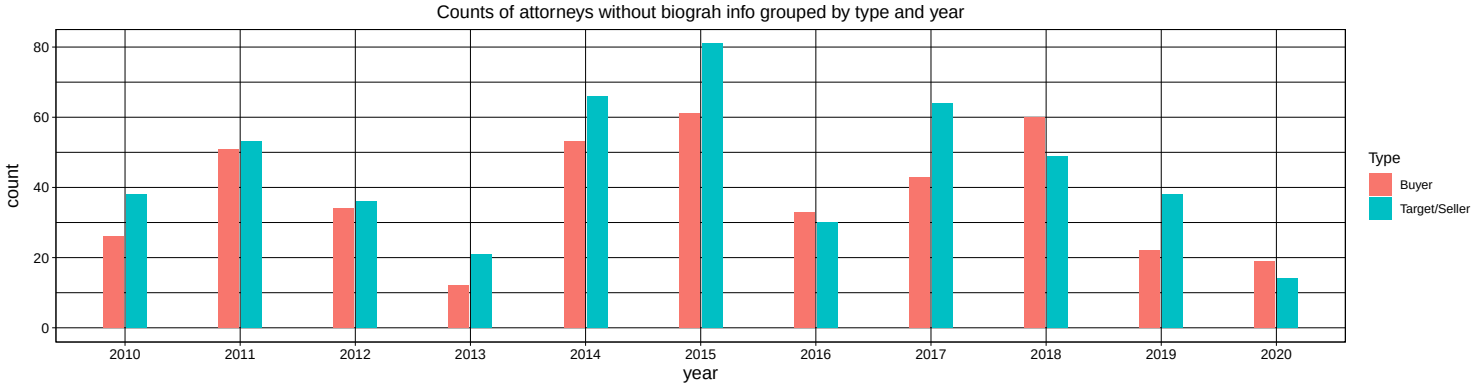
```
> three_matched_stacked <- read_csv("../data/confidence_match/three_matched_stacked.csv",
+     col_types = cols(.default = "c"))
> # View(three_matched_stacked)
>
> # If duplicates, keep attorneys from MA only
> dis_three_matched_stacked <- three_matched_stacked %>%
+     distinct(Deal_number, `Signing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





How many deals are missing attorneys' biographical information?

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

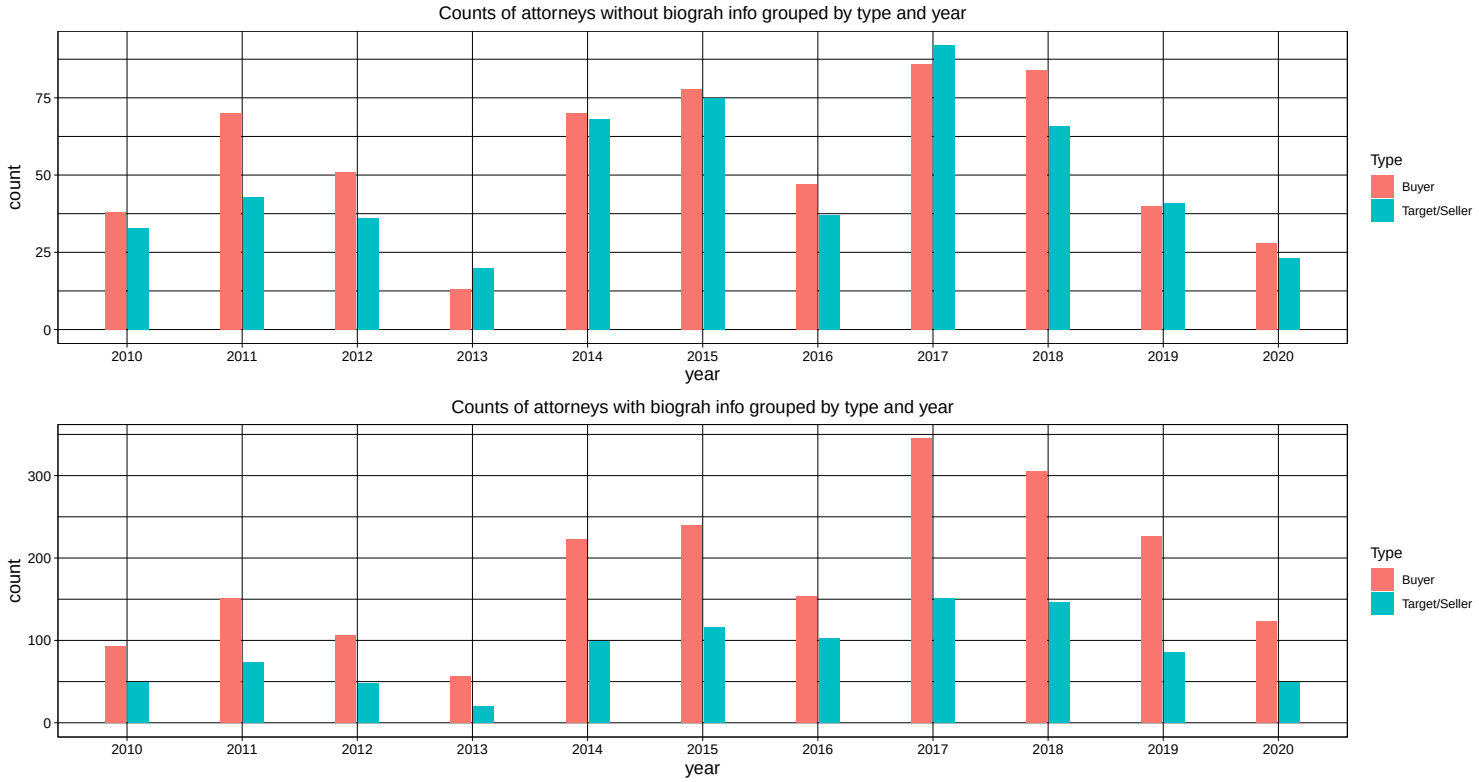
```
## [1] 1139
```

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 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

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 3: 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 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.

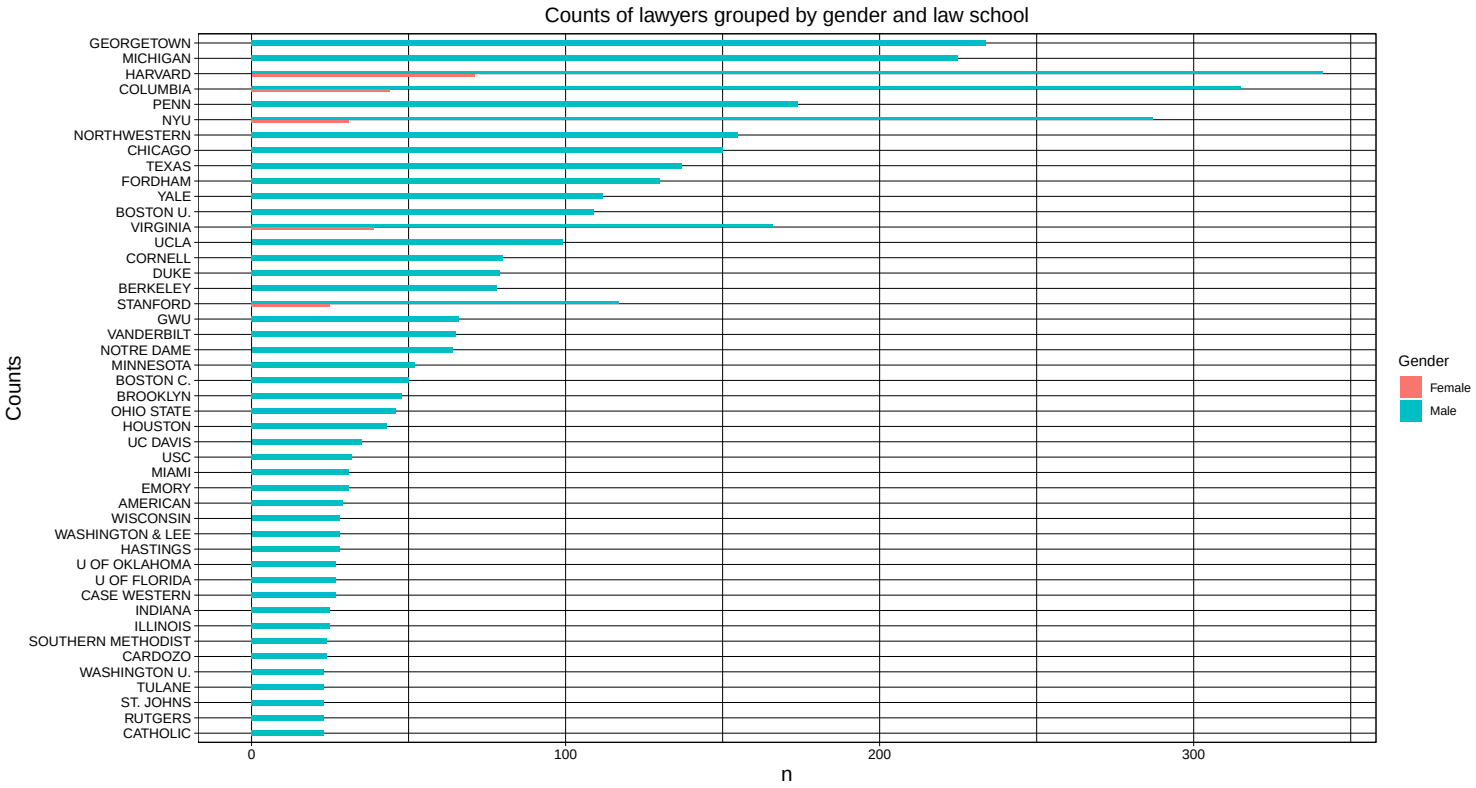
```
> dis_three_matched_stacked %>%
+     drop_na(Gender, `Law School`) %>%
+     group_by(`Law School`, Law_Firm) %>%
+     count() %>%
+     arrange(desc(n)) %>%
+     write_csv(file = "../data/deal_lawyer/distri_law_sch_firm.csv")
```

Table 4: 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

```
> 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 a
+     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
```



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"))
```

```
> 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"))
```

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 an
```

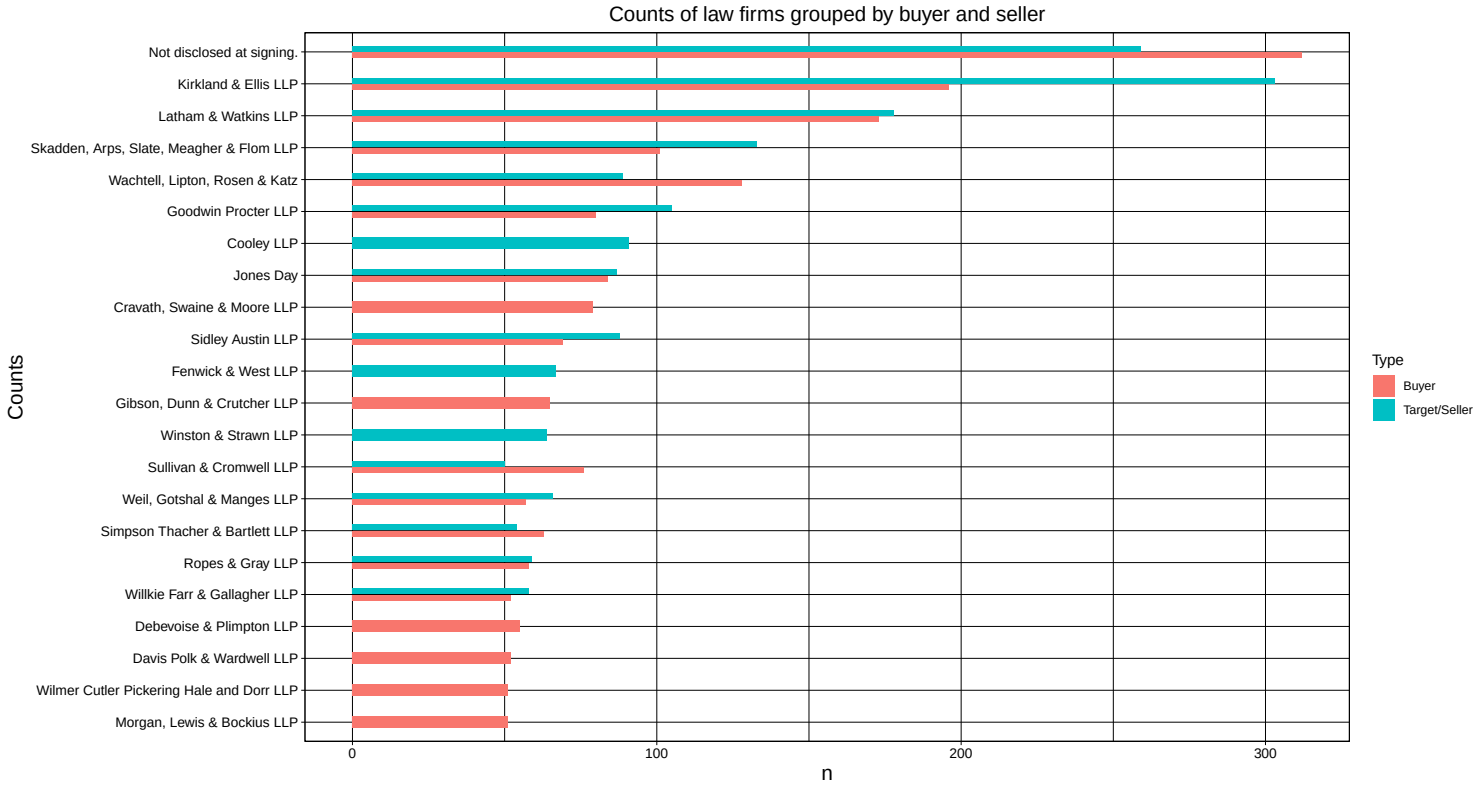
Table 5: Industry Sector Count

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

Table 6: 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

```
+     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 7: MA Lawyers grouped by Buyer/Seller

Type	n
Target/Seller	2685
Buyer	2660

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.

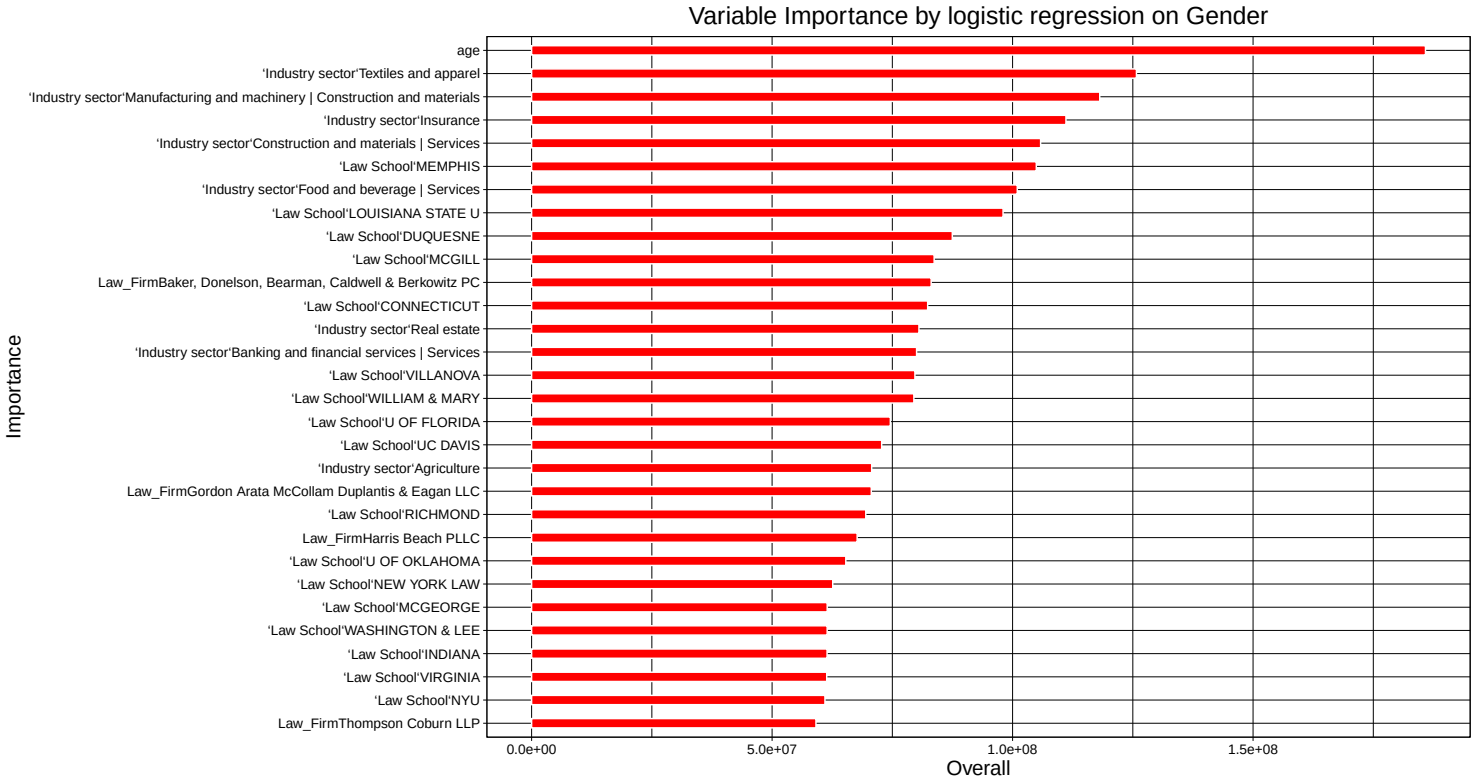
First we used Multiple Logistic Regression with all variables

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

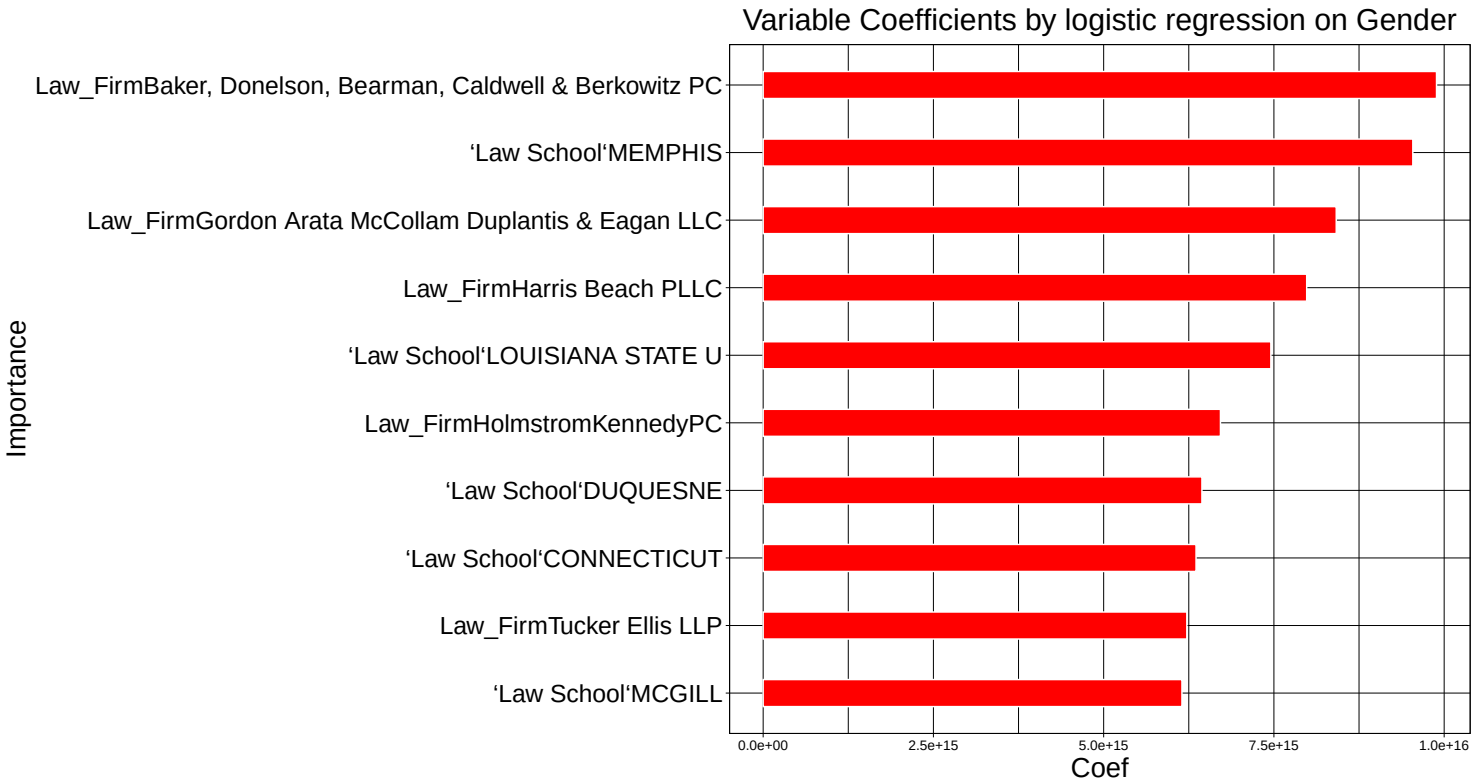
```
## Warning: glm.fit: algorithm did not converge
```

```
## 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

```
## [1] 0
```

```
## [1] 0
```

Law_Firm coefficient signs

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

```
## [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
```

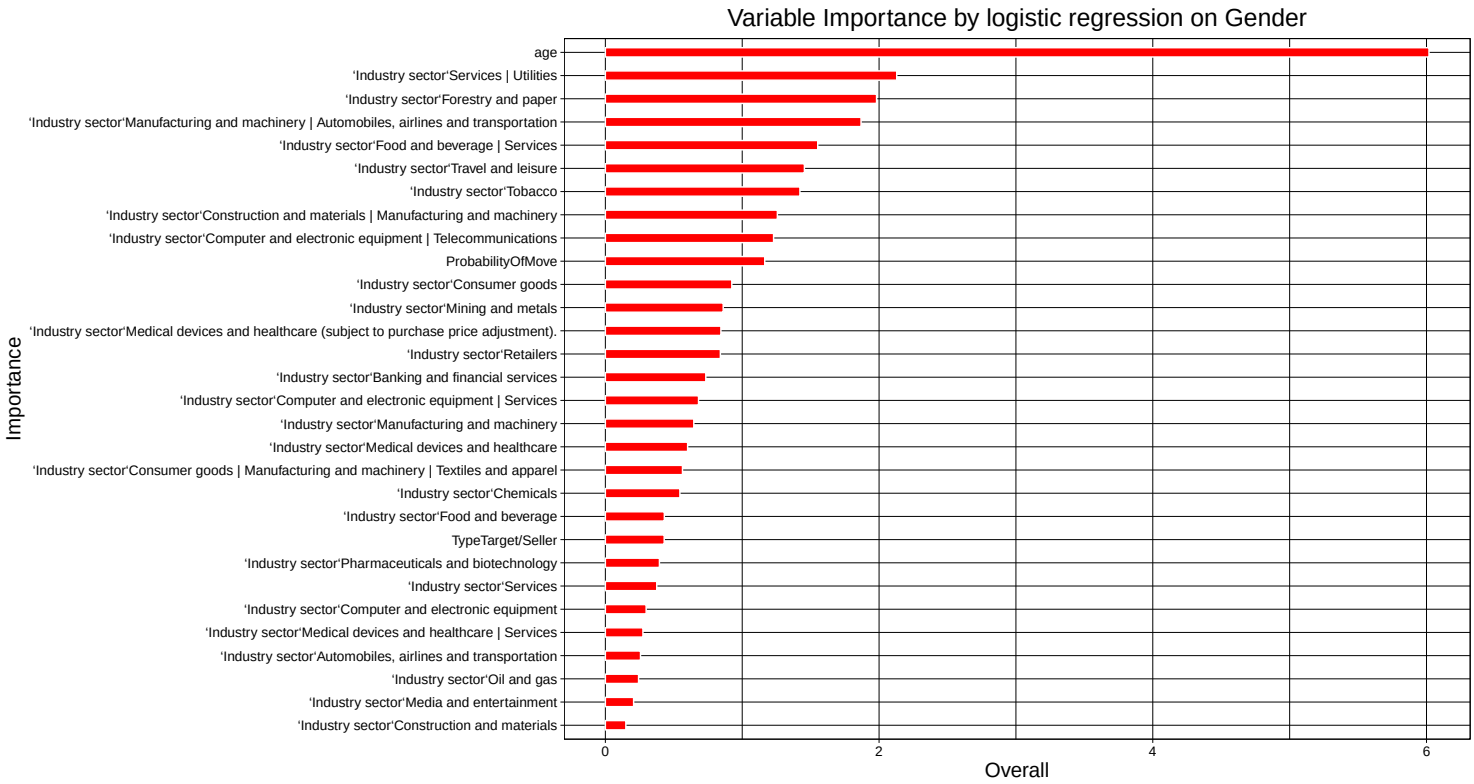
```
> data.frame(Coef = logit$coefficients) %>%
+     rownames_to_column(var = "variable") %>%
+     filter(grepl("Industry", variable)) %>%
+     filter(Coef < 0) %>%
+     nrow
```

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

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

```
> logit <- glm(Gender ~ . - Law_Firm - `Law School`, data = dat, family = "binomial")
> varImp(logit) %>%
+     arrange(desc(Overall)) %>%
+     rownames_to_column(var = "variable") %>%
+     top_n(30) %>%
+     ggplot(aes(y = reorder(variable, Overall), x = Overall)) + geom_col(position = position_dodge2(),
+     width = 0.4, orientation = "y", fill = "red", color = "white") + labs(title = "Variable Importance by log
+     y = "Importance") + 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 = 18))
```

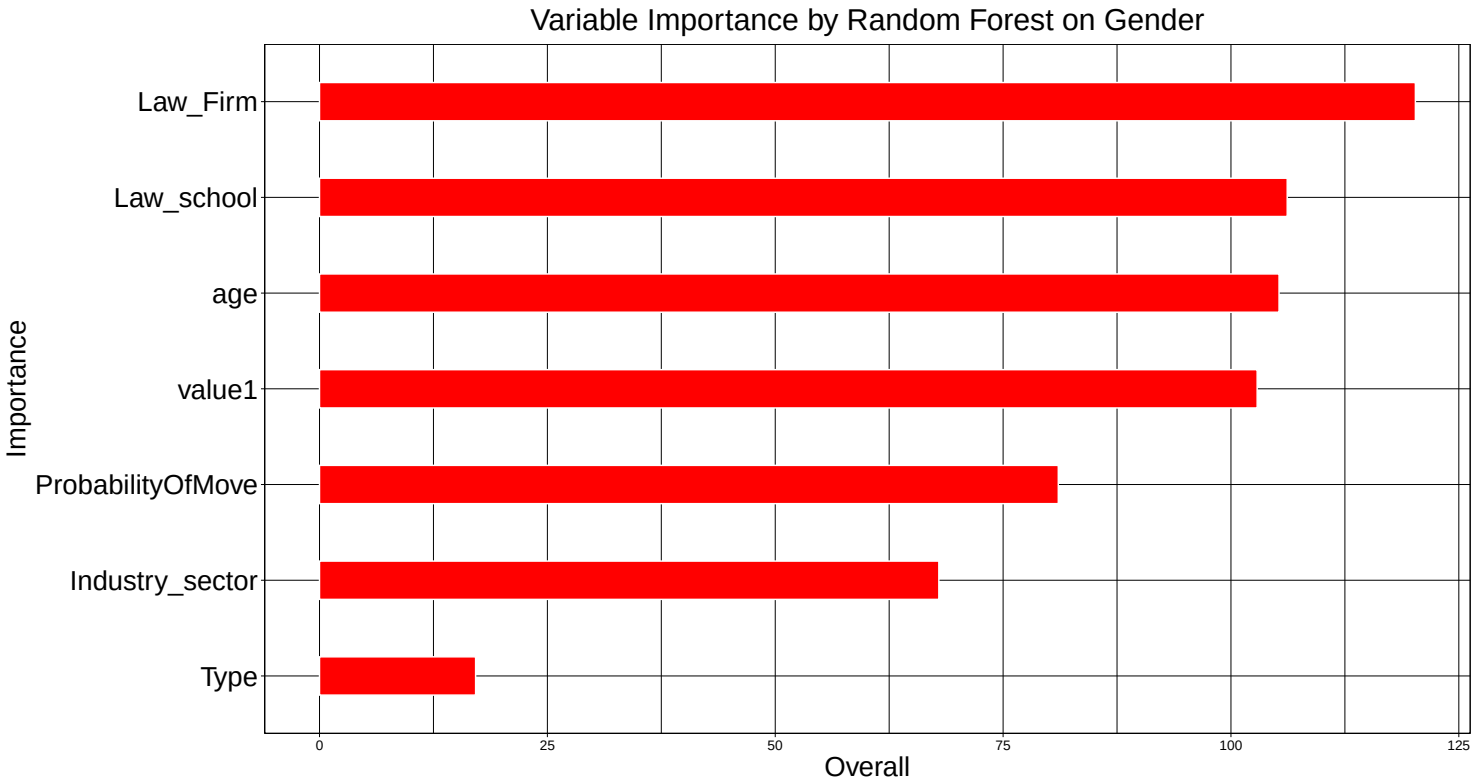
```
## Selecting by Overall
```



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



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.