Indeed Reviews: Descriptives

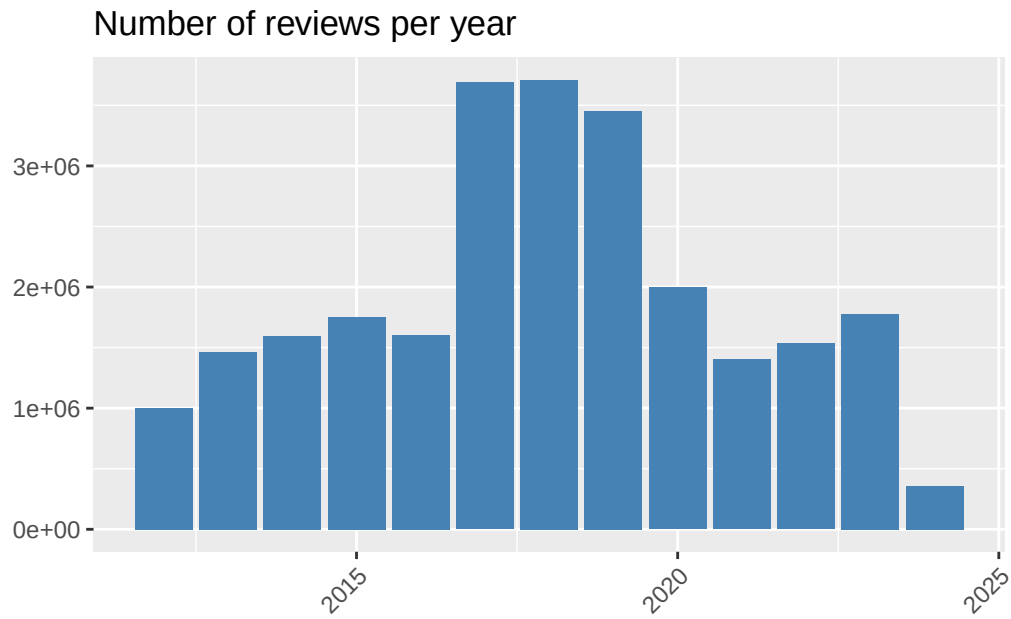
Set up

```
library(tidyverse)
library(lubridate)
library(stringi)
library(knitr)


d <- read_csv("../data/indeed_reviews_clean.csv", locale = locale(encoding = "Latin1")) |>
  select(-...1)
```

Descriptives

```
d |>
  ggplot(aes(x = year)) +
  geom_bar(fill = "steelblue") +
  labs(x = "", y = "",
       title = "Number of reviews per year") +
  theme(axis.text.x = element_text(angle = 45,
                                   hjust = 1,
                                   size = 9))
```



Missingness

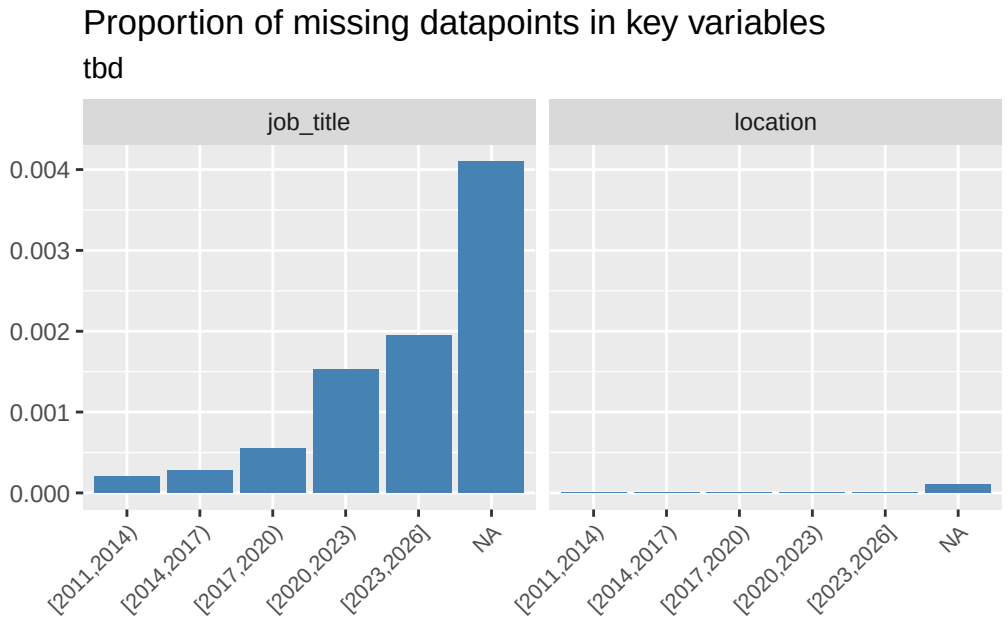
Job data

```
miss_d <- d |>
  select(yr3_int, job_title, location) |>
  mutate(across(c(job_title, location), ~is.na(.), .names = "miss_{.col}")) |>
  pivot_longer(cols = starts_with("miss_"),
               names_to = "var",
               values_to = "missing") |>
  mutate(var = sub('^miss_', "", var)) |>
  group_by(yr3_int, var) |>
  summarise(total = n(),
            n_missing = sum(missing),
            prop_missing = mean(missing))

miss_d |> kable()
```

yr3_int	var	total	n_missing	prop_missing
[2011,2014)	job_title	2465179	508	0.0002061
[2011,2014)	location	2465179	0	0.0000000
[2014,2017)	job_title	4953354	1392	0.0002810
[2014,2017)	location	4953354	0	0.0000000
[2017,2020)	job_title	10853113	5935	0.0005468
[2017,2020)	location	10853113	2	0.0000002
[2020,2023)	job_title	4929729	7552	0.0015319
[2020,2023)	location	4929729	1	0.0000002
[2023,2026]	job_title	2133276	4146	0.0019435
[2023,2026]	location	2133276	4	0.0000019
NA	job_title	339603	1394	0.0041048
NA	location	339603	37	0.0001090

```
miss_d |>
  ggplot(aes(x = yr3_int, y = prop_missing)) +
  geom_col(fill = "steelblue") +
  facet_wrap(~var) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +
  labs(x = "", y = "",
       title = "Proportion of missing datapoints in key variables",
       subtitle = "tbd")
```



Date

```
d |>
  filter(is.na(date)) |>
  tally() |>
  knitr::kable()
```

n
40

Reviews

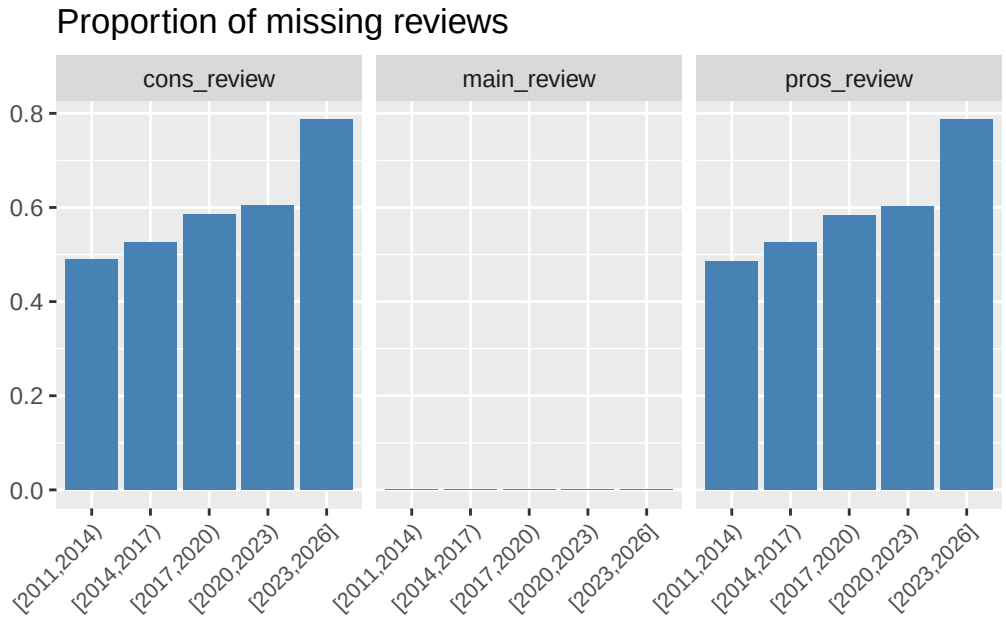
```
vars <- c("main_review", "pros_review", "cons_review")
miss_d <- d |>
  select(yr3_int, main_review, pros_review, cons_review) |>
  filter(!is.na(yr3_int)) |>
  mutate(across(all_of(vars), ~is.na(.), .names = "miss_{.col}")) |>
  pivot_longer(cols = starts_with("miss_"),
               names_to = "var",
               values_to = "missing") |>
  mutate(var = sub('^miss_', "", var)) |>
  group_by(yr3_int, var) |>
  summarise(total = n(),
            n_missing = sum(missing),
            prop_missing = mean(missing))

miss_d |> kable()
```

yr3_int	var	total	n_missing	prop_missing
[2011,2014)	cons_review	2465179	1207246	0.4897194
[2011,2014)	main_review	2465179	0	0.0000000
[2011,2014)	pros_review	2465179	1198196	0.4860483
[2014,2017)	cons_review	4953354	2608326	0.5265777
[2014,2017)	main_review	4953354	0	0.0000000
[2014,2017)	pros_review	4953354	2600322	0.5249619
[2017,2020)	cons_review	10853113	6345829	0.5847013
[2017,2020)	main_review	10853113	0	0.0000000

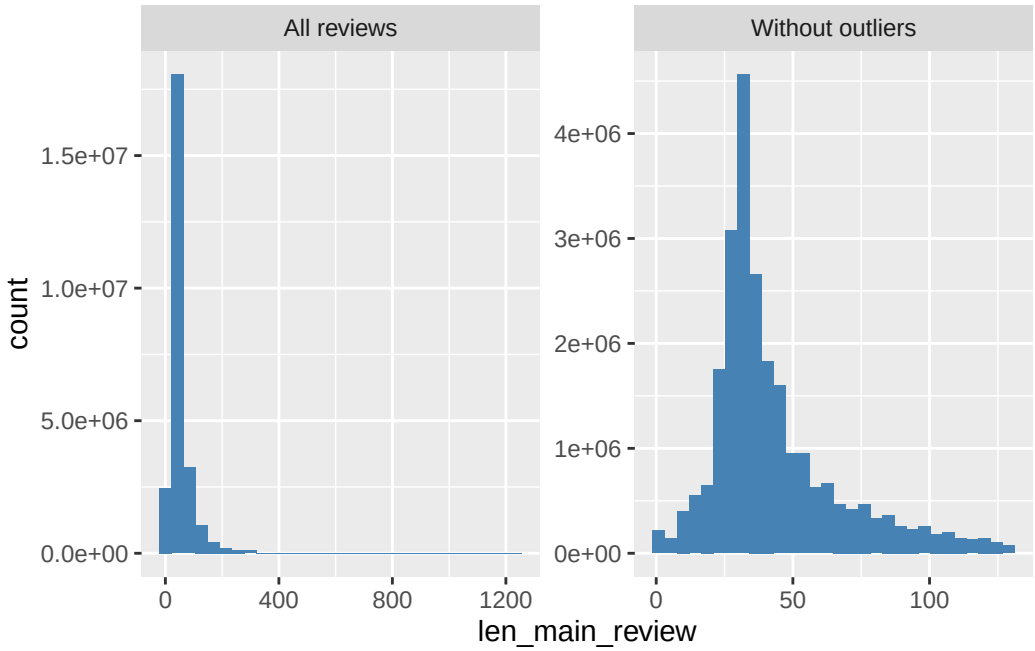
yr3_int	var	total	n_missing	prop_missing
[2017,2020)	pros_review	10853113	6337762	0.5839580
[2020,2023)	cons_review	4929729	2974788	0.6034384
[2020,2023)	main_review	4929729	0	0.0000000
[2020,2023)	pros_review	4929729	2973675	0.6032127
[2023,2026]	cons_review	2133276	1677437	0.7863197
[2023,2026]	main_review	2133276	0	0.0000000
[2023,2026]	pros_review	2133276	1677134	0.7861777

```
miss_d |>
  ggplot(aes(x = yr3_int, y = prop_missing)) +
  geom_col(fill = "steelblue") +
  facet_wrap(~var) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +
  labs(x = "", y = "", title = "Proportion of missing reviews")
```



Text lengths

```
bind_rows(d |> mutate(subset = "All reviews"),
          d |> filter(len_main_review > 0 &
                len_main_review < (mean(len_main_review) + 2*sd(len_main_review))) |>
    mutate(subset = "Without outliers")) |>
  ggplot(aes(x = len_main_review)) +
  geom_histogram(fill="steelblue") +
  facet_wrap(~subset, scales = "free")
```




```
len_d <- d |>
  filter(!is.na(yr3_int),
         len_main_review < (mean(len_main_review) + 2*sd(len_main_review))) |>
  group_by(yr3_int) |>
  summarise(avg_len = mean(len_main_review),
            sd = sd(len_main_review),
            min = min(len_main_review),
            max = max(len_main_review),
            n = n(),
            .groups = "drop")

len_d |> kable()
```

yr3_int	avg_len	sd	min	max	n
[2011,2014)	39.62924	28.75808	1	129	2302077
[2014,2017)	41.91693	25.70459	1	129	4660390
[2017,2020)	42.97340	20.37891	1	129	10486473
[2020,2023)	46.13093	22.47183	1	129	4658376
[2023,2026]	49.63475	25.96247	1	129	1989132

```
len_d |>
  ggplot(aes(y=avg_len, x=yr3_int)) +
  geom_col(fill="steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +
  labs(x = "", y = "",
       title = "Average review length in 3 year periods",
       subtitle = "Ignoring ourliers")
```