Webinar: The real life of a data analyst. Gender gap in Kazakh wages using R

Data analyst's toolkit. How to analyze public data and stay objective.

Elvira Nassirova 

Thanks for coming!

- 4 years experience as Data Analyst at post office, telecom, bank, classified and product studio
- Product Data Analyst at Railsware
- Mentor at Yandex.Practicum, Data Analysis program

LinkedIn https://www.linkedin.com/in/nassirova/

Data Analyst toolkit

Example tasks

- answer question "Why?"
- get some data
- automate something
- data monitoring
- explain your findings
- learn something new everyday
- collaborate

Toolkit

SQL

- Data extraction
- Data preparation
- Data monitoring

Toolkit

R / Python + Google Colab / Jupyter Notebook

- Exploratory Data Analysis
- One-time deep reports, especially including text analytics
- API integrations and crawling

Toolkit

Dashboards, dataviz

- flexdashboard
- Dash
- Data Studio (free)
- Power BI
- Tableau

Toolkit

Automation

- Google Apps Scripts
- Google Cloud functions + Python / js
- Terraform
- Clasp

Toolkit

GitHub

- Version control
- Deployment via Git Actions

Questions

How to analyze public data and stay objective?

A bit of context

Two points of view

- everything is equal between men and women **vs.** not at all

Can I use numbers to answer that dilemma?

- How big is a pay gap between men and women in Kazakhstan and why?

Datasource (at first)

- Public data from statistic department of Kazakhstan

Links

- GitHub repo
- article on Medium (ru)

The Code

```r
# reading table data from MS Word

salary_location <- './data/salary.docx'
salary_doc <- docxtractr::read_docx(salary_location)

# We need 70th table from the doc (don't ask me how I know this :D)
salary_industry <- docx_extract_tbl(salary_doc, 70)
```

```r
# data preparation tricks

salary_industry <-
  docx_extract_tbl(salary_doc, 70) %>%
  setNames(., c('field_kz', 'men', 'women', 'field_ru')) %>%

# use mutate_at(vars()) to apply chanes for several columns
  mutate_at(vars(contains('men')), ~salary_to_num(.)) %>%

# use everyting() to change order for only certain fields
  select(field_ru, everything()) %>%

# use gather to unpivot data
  tidyr::gather(sex, salary, -field_ru) %>%

# use recode for new names in character variables
  mutate(sex = recode(sex, 'men' = 'Мужчины', 'women' = 'Женщины'),

# user reorder to order factor variable by integer variable
         sex = reorder(sex, salary))
```

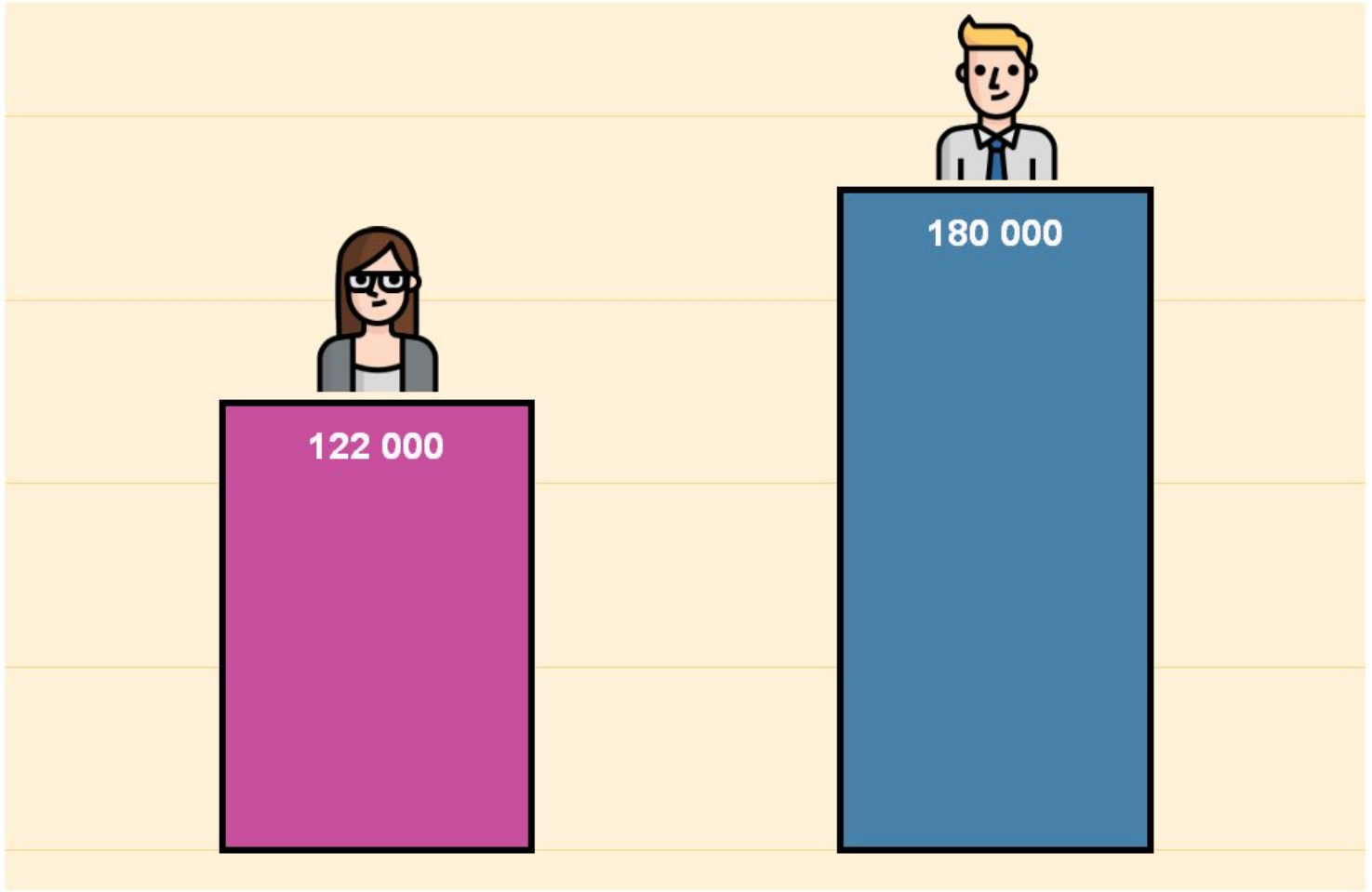
```r
# Add icons to ggplot2 plot

# icon for woman
woman_banker <-
  magick::image_read_svg('./icons/woman.svg', width = 150) %>%
  grid::rasterGrob(., interpolate = T)

# icon for man
man_banker <-
  magick::image_read_svg('./icons/man.svg', width = 150) %>%
  grid::rasterGrob(., interpolate = T)

# chart stuff
salary_dynamics %>%
  filter(year == 2017) %>%
  mutate(salary = round(salary / 1000) * 1000,
         salary_label = scales::number(salary)) %>%
  ggplot(aes(x = sex, y = salary, fill = sex)) +
  geom_bar(stat = 'identity', position = 'dodge', width = .5, color = 'black', size = 1.5) +
  geom_text(aes(label = salary_label), vjust = 2, size = 6, fontface = 'bold', col = '#FFFAFF') +
  scale_y_continuous(label = k_formatter, limits = c(0, 220000)) +
  scale_fill_manual(values = sex_colors_dark) +
  labs(x = '', y = '') +

# adding icons
  annotation_custom(woman_banker, ymin = 125000, ymax = 170000, xmin = .8, xmax = 1.2) +
  annotation_custom(man_banker, ymin = 182782, ymax = 227782, xmin = 1.8, xmax = 2.2) +
  viz_theme
```


122 000
180 000

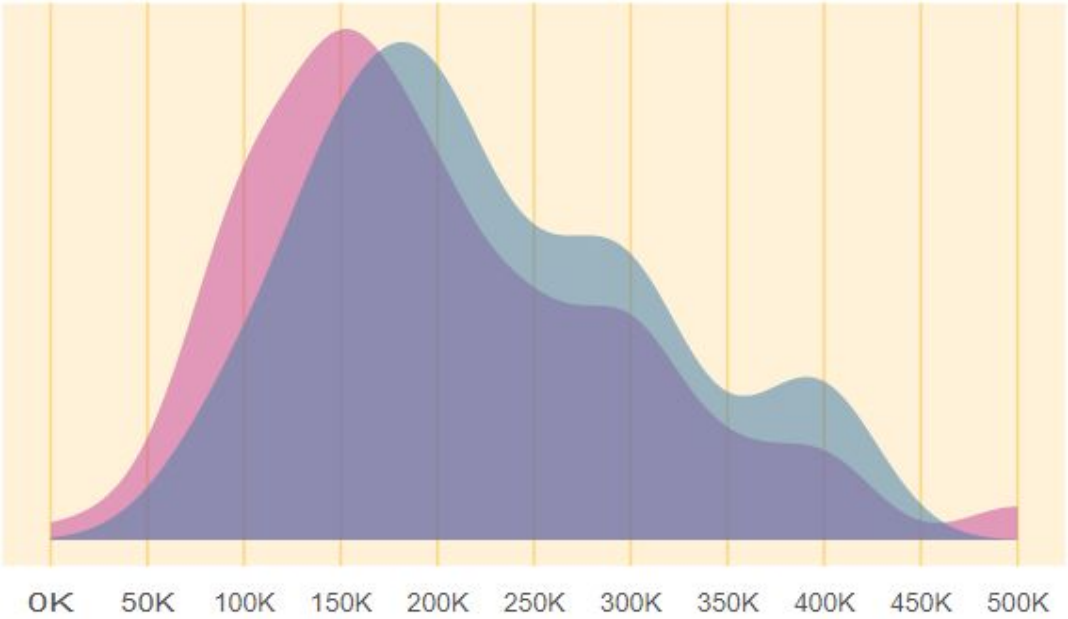
32% pay gap

but why?

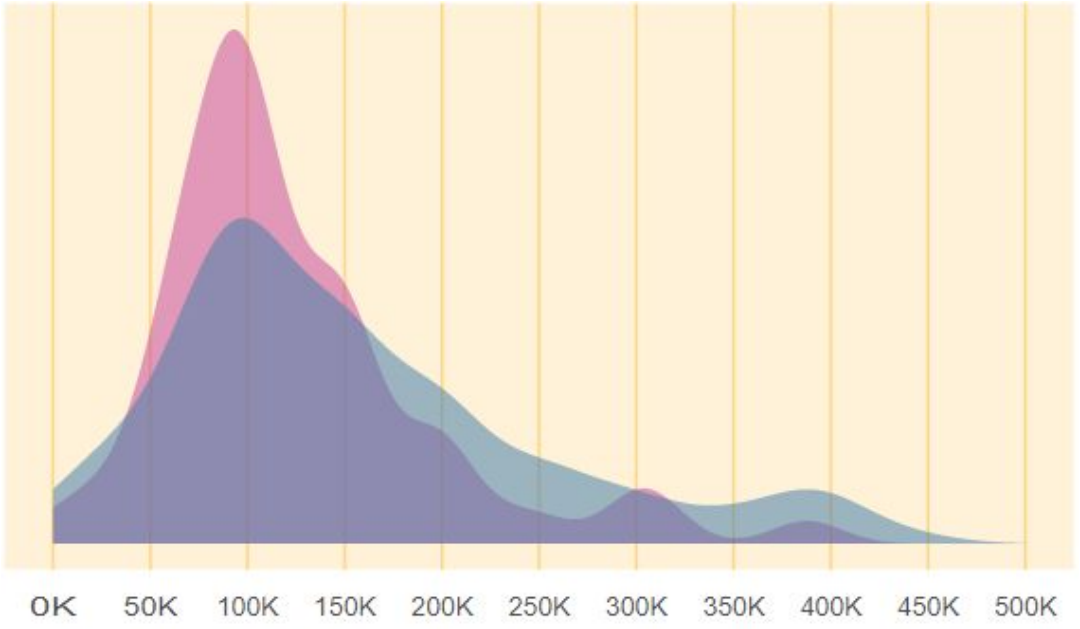
Do women ask for less money?

I decided to analyze CVs on HeadHunter being one of the most popular jobs classified and compare asked salary

Financials



FrontEnd Developers



Based on that statistics — yes

But “why” is another deep question :)

Questions