FGV MBA - Trabalho de Análise Exploratória de Dados

Daniel Campos, Leandro Daniel, Rodrigo Goncalves e Ygor Lima

2019-06-30

1 - Berka Bank (setting the scene)

1.1 - Domain

Onde upon a time, there was a bank offering services to private persons. The services include managing of accounts, offerings loans, etc.

1.2 - Task descripion

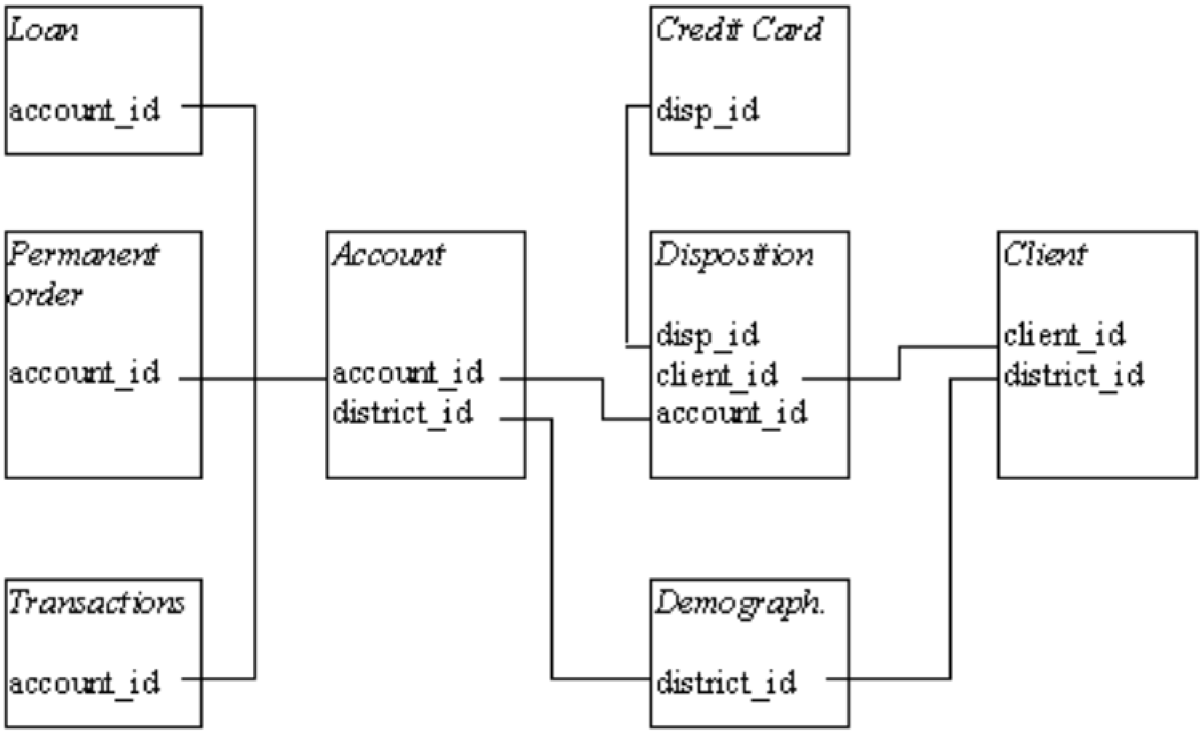
The bank wants to improve their services. For instance, the bank managers have only vague idea, who is a good client (whom to offer some additional services) and who is a bad client (whom to watch carefully to monimize the bank loses).

Fortunately, the bank stores data about their clients, the accounts (transactions within several months), the loans already granted, the credit cards issued.

The bank managers hope to improve their undestanding of customers and seed specific actions to improve services.

A mere application of discovery tool will not be convincing for them.

1.3 - Data description



1.4 - Project at GitHub

This project can be found and downloaded at GitHub: https://github.com/ldaniel/R__Bank__Berka

> Valar Morghulis! :)

---

2 - Data ingestion, cleaning, translation and enhancement

Before starting the Berka Analysis, a few important steps were taken in order to prepare the source data files. These steps are listed below:

- **Step 01**: Create Functions
- **Step 02**: Data Ingestion
- **Step 03**: Data Cleaning
- **Step 04**: Label Translation
- **Step 05**: Data Enhancement

2.1 - Create Functions

To-do.

2.2 - Data Ingestion

To-do.

2.3 - Data Cleaning

```
sapply(transaction, function(x) sum(is.na(x)))
```

```
sapply(permanent_order, function(x) table(as.character(x) =="")["TRUE"])
```

2.4 - Label Translation

To-do.

2.5 - Data Enhancement

To-do.

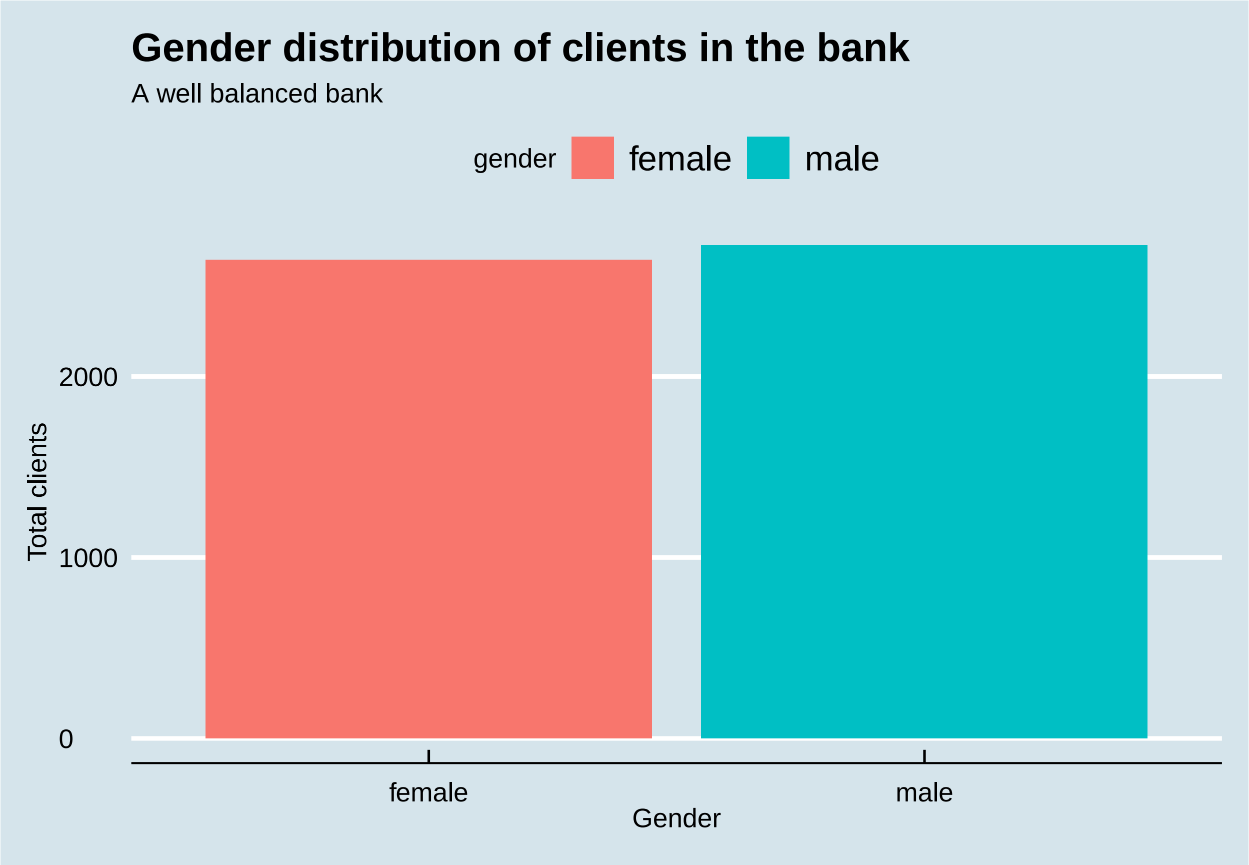
---

3 - The Berka Bank Analysis

3.1 - Gender Exploration

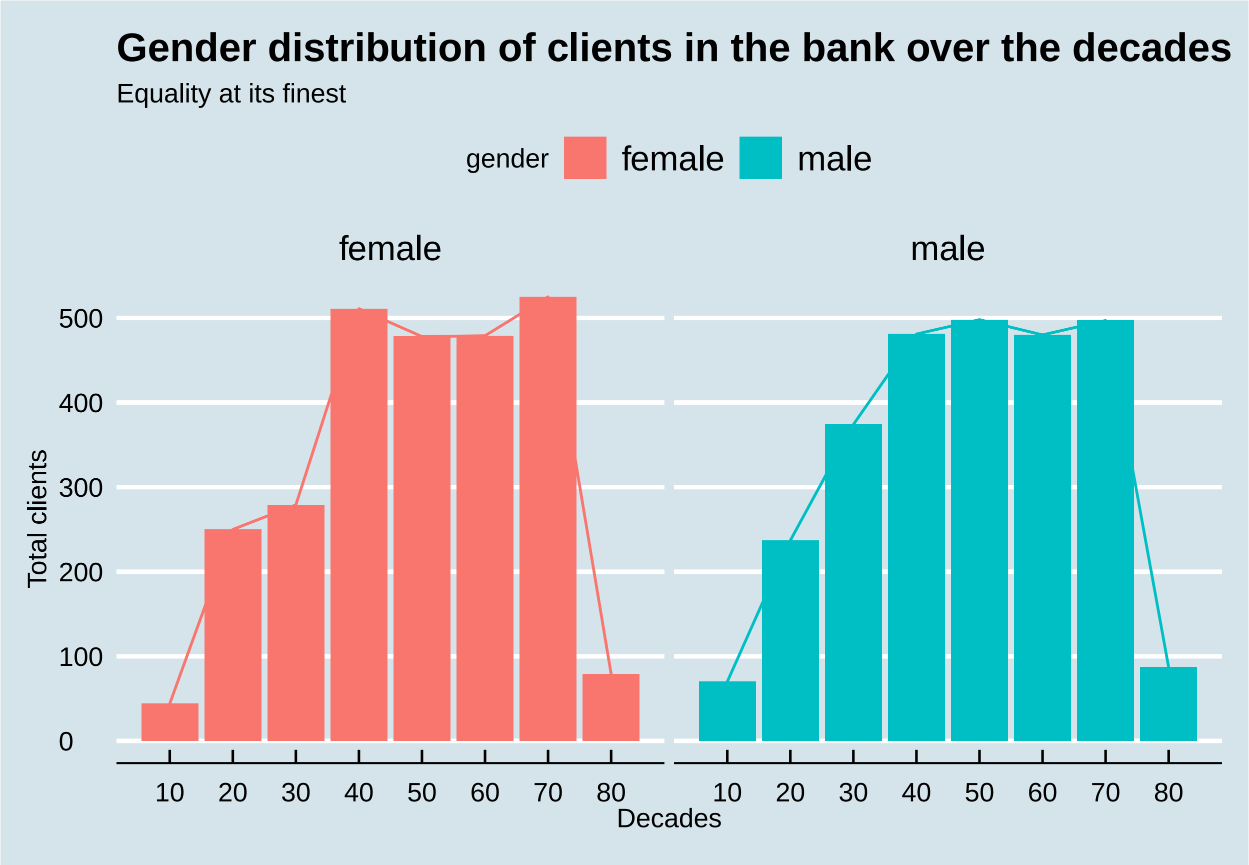
At first glance, gender equality is well balanced in the bank, even when observed over the decades. Even more impressive, gender equality is everywhere in the country.

```r
# gender distribution of clients in the bank
ggplot(data = client) +
  aes(x = gender, fill = gender) +
  geom_bar() +
  labs(title = "Gender distribution of clients in the bank",
       subtitle = "A well balanced bank",
       x = "Gender",
       y = "Total clients") +
  theme_economist()
```



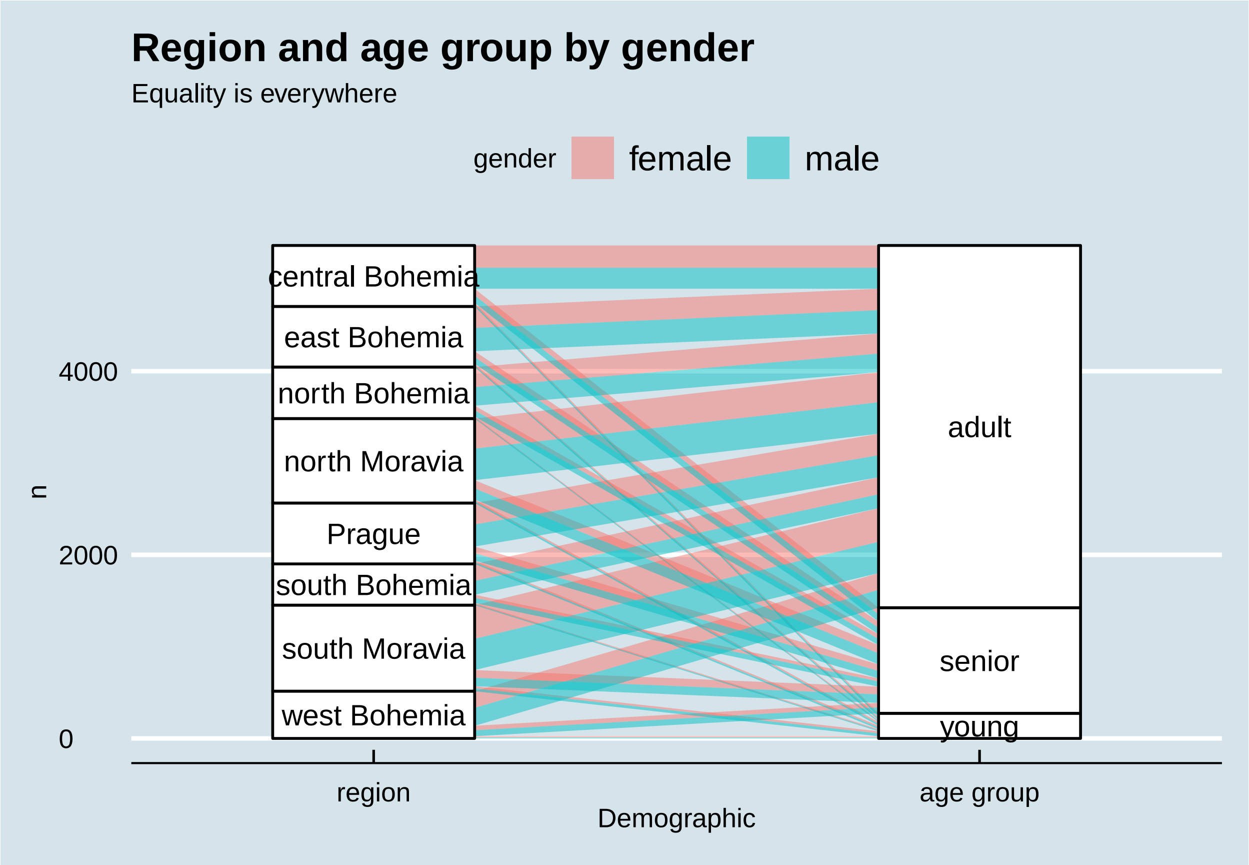
```r
clientGenderOverDecades <- client %>%
  group_by(decade = as.integer(substr(client$birth_number, 1,1)) * 10, gender = client$gender) %>%
  count()

# gender distribution of clients in the bank over the decades
ggplot(data = clientGenderOverDecades) +
  aes(x = decade, fill = gender, weight = n) +
  scale_x_continuous(breaks = c(0, 10, 20, 30, 40, 50, 60, 70, 80)) +
  geom_bar() +
  geom_line(aes(y = n, color = gender)) +
  labs(title = "Gender distribution of clients in the bank over the decades",
       subtitle = "Equality at its finest",
       x = "Decades",
       y = "Total clients") +
  theme_economist() +
  facet_wrap(vars(gender))
```



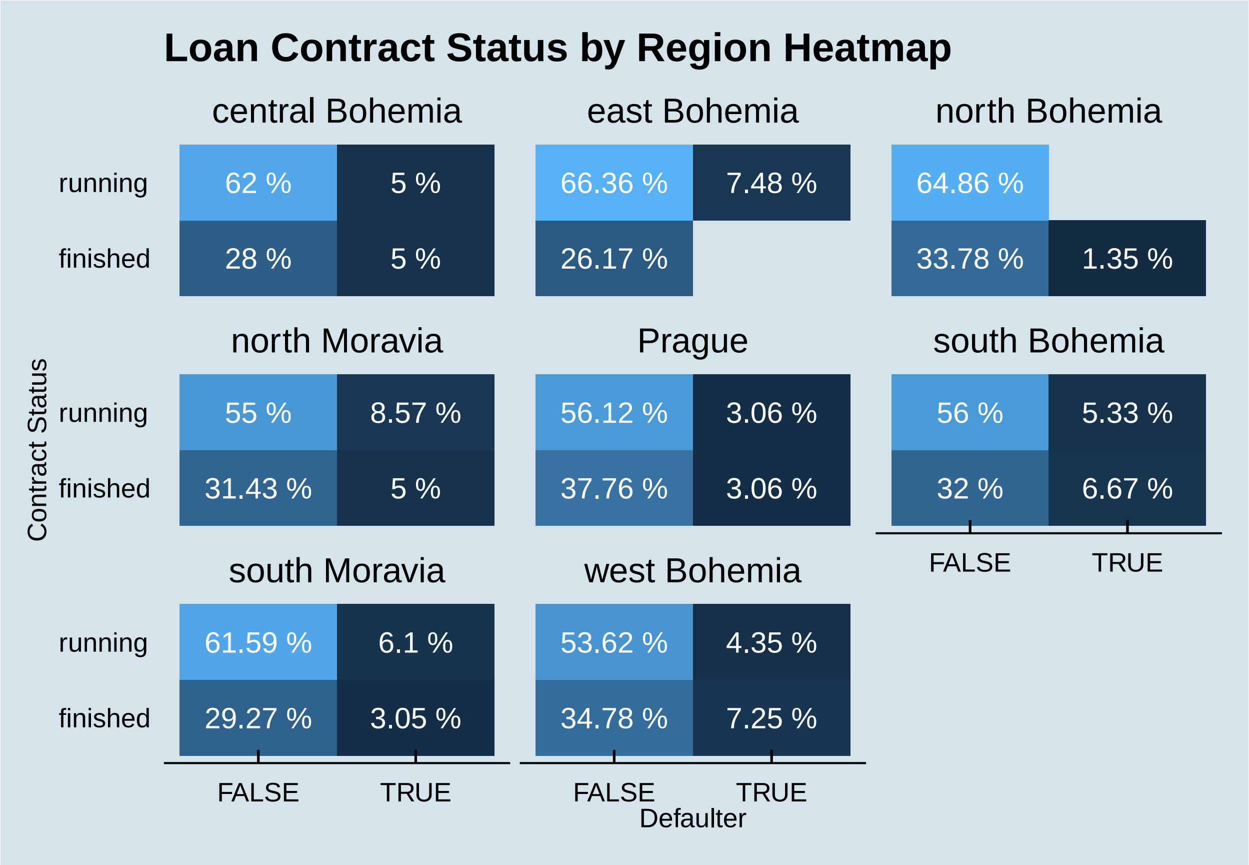
```r
# alluvial diagram representation of gender, age group and region
clientGenderAgeGroupByRegion <- client %>%
  mutate(age_group = ifelse(age < 21, "young", ifelse(age >= 21 & age <= 60, "adult", "senior"))) %>%
  inner_join(district, by = "district_id") %>%
  group_by(age_group, gender, region) %>%
  count()

ggplot(data = clientGenderAgeGroupByRegion, aes(axis1 = region, axis2 = age_group, y = n)) +
  scale_x_discrete(limits = c("region", "age group"), expand = c(.1, .1)) +
  xlab("Demographic") +
  geom_alluvium(aes(fill = gender), knot.pos = 0) +
  geom_stratum() +
  geom_text(stat = "stratum", label.strata = TRUE) +
  theme_economist() +
  ggtitle("Region and age group by gender", "Equality is everywhere")
```

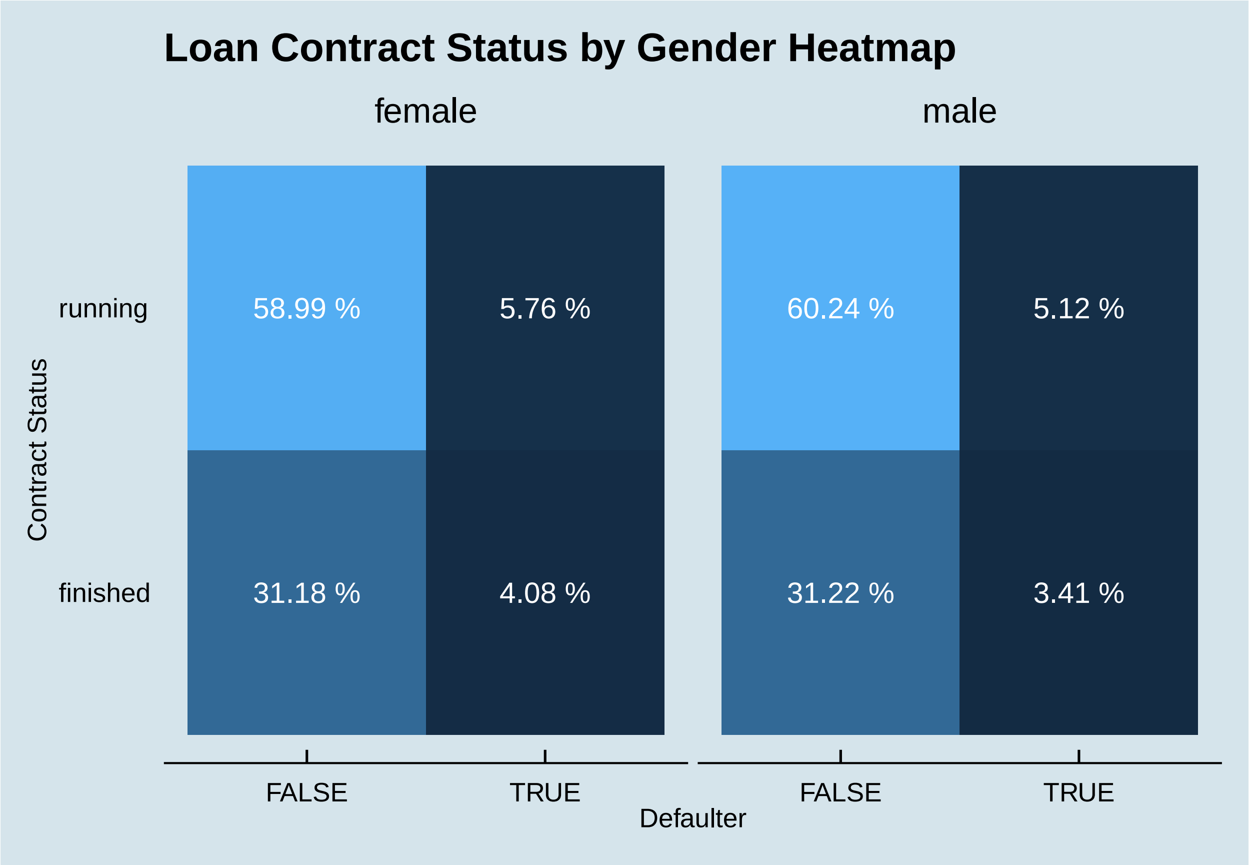


3.2 - Loan Exploration

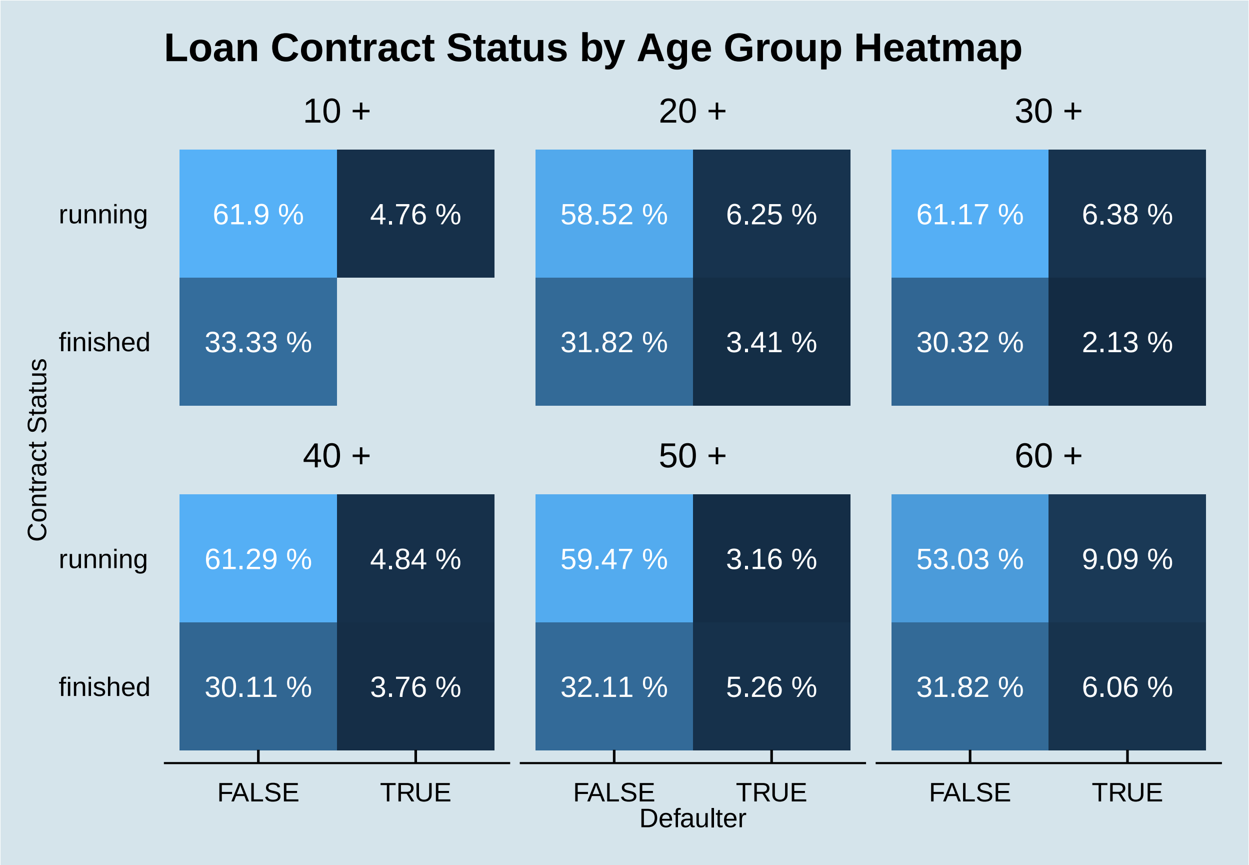
```
left_join(loan, disposition, by = 'account_id') %>%
  left_join(client, by = 'client_id') %>%
  left_join(district, by = 'district_id') %>%
  group_by(region, contract_status, defaulter) %>%
  summarise(count = n(),
            amount = sum(amount)) %>%
  group_by(region, contract_status) %>%
  mutate(count_contract_status = sum(count),
         amount_contract_status = sum(amount)) %>%
  group_by(region) %>%
  mutate(count_region = sum(count),
         amount_region = sum(amount)) %>%
  ggplot(aes(x = defaulter, y = contract_status, fill = count / count_region)) +
  geom_bin2d(stat = 'identity') +
  geom_text(aes(label = paste(round(count / count_region * 100, 2), '%')), color = 'white') +
  facet_wrap(~region) +
  theme_economist() +
  theme(legend.position = 'none', panel.grid.major = element_blank(), panel.grid.minor = element_blank(
  labs(x = 'Defaulter', y = 'Contract Status', title = 'Loan Contract Status by Region Heatmap')
```



```
left_join(loan, disposition, by = 'account_id') %>%
  left_join(client, by = 'client_id') %>%
  left_join(district, by = 'district_id') %>%
  group_by(gender, contract_status, defaulter) %>%
  summarise(count = n(),
            amount = sum(amount)) %>%
  group_by(gender, contract_status) %>%
  mutate(count_contract_status = sum(count),
         amount_contract_status = sum(amount)) %>%
  group_by(gender) %>%
  mutate(count_gender = sum(count),
         amount_gender = sum(amount)) %>%
  ggplot(aes(x = defaulter, y = contract_status, fill = count / count_gender)) +
    geom_bin2d(stat = 'identity') +
    geom_text(aes(label = paste(round(count / count_gender * 100, 2), '%')), color = 'white') +
    facet_wrap(~gender) +
    theme_economist() +
    theme(legend.position = 'none', panel.grid.major = element_blank(), panel.grid.minor = element_blan
    labs(x = 'Defaulter',
         y = 'Contract Status',
         title = 'Loan Contract Status by Gender Heatmap')
```



```r
left_join(loan, disposition, by = 'account_id') %>%
  left_join(client, by = 'client_id') %>%
  left_join(district, by = 'district_id') %>%
  group_by(age_bin, contract_status, defaulter) %>%
  summarise(count = n(),
            amount = sum(amount)) %>%
  group_by(age_bin, contract_status) %>%
  mutate(count_contract_status = sum(count),
         amount_contract_status = sum(amount)) %>%
  group_by(age_bin) %>%
  mutate(count_age_bin = sum(count),
         amount_age_bin = sum(amount)) %>%
  ggplot(aes(x = defaulter, y = contract_status, fill = count / count_age_bin)) +
    geom_bin2d(stat = 'identity') +
    geom_text(aes(label = paste(round(count / count_age_bin * 100, 2), '%')), color = 'white') +
    facet_wrap(~age_bin) +
    theme_economist() +
    theme(legend.position = 'none', panel.grid.major = element_blank(), panel.grid.minor = element_blan
    labs(x = 'Defaulter',
         y = 'Contract Status',
         title = 'Loan Contract Status by Age Group Heatmap')
```

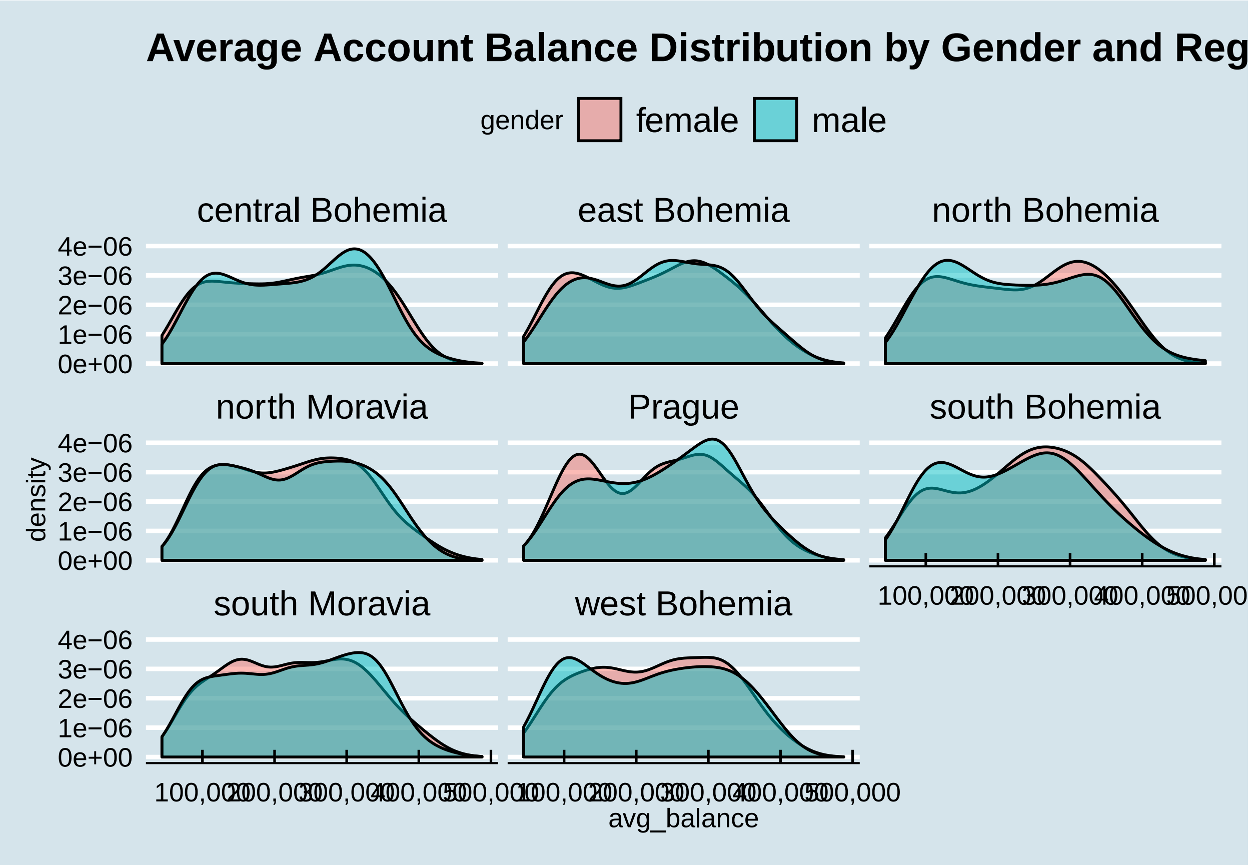


3.3 - Account Balance Exploration

```
account_balance <- arrange(transaction, desc(date), account_id) %>%
  group_by(account_id) %>%
  mutate(avg_balance = mean(balance)) %>%
  filter(row_number() == 1) %>%
  select(account_id, date, balance, avg_balance)

colnames(account_balance) <- c("account_id", "last_transaction_date", 'account_balance', 'avg_balance')

left_join(account_balance, disposition, by = 'account_id') %>%
  left_join(client, by = 'client_id') %>%
  left_join(district, by = 'district_id') %>%
  filter(type == 'Owner') %>%
  ggplot(aes(avg_balance)) +
    geom_density(alpha = 0.5, aes(fill = gender)) +
    scale_x_continuous(labels = scales::comma) +
    labs(title = 'Average Account Balance Distribution by Gender and Region') +
    theme_economist() +
    facet_wrap(~region)
```


Average Account Balance Distribution by Gender and Reg
gender
female
male
central Bohemia
east Bohemia
north Bohemia
north Moravia
Prague
south Bohemia
south Moravia
west Bohemia
density
4e−06
3e−06
2e−06
1e−06
0e+00
100,000
200,000
300,000
400,000
500,000
avg_balance