Model building

Cross validation

Prof. Dr. Jan Kirenz

IMDB ratings

Code Plot

```
ggplot(office_ratings, aes(x = imdb_rating)) +
  geom_histogram(binwidth = 0.25) +
  labs(
    title = "The Office ratings",
    x = "IMDB Rating"
  )
```

IMDB ratings vs. number of votes

Code Plot

```
ggplot(office_ratings, aes(x = total_votes, y = imdb_rating, color = season)) +
  geom_jitter(alpha = 0.7) +
  labs(
    title = "The Office ratings",
    x = "Total votes",
    y = "IMDB Rating",
    color = "Season"
  )
```

Outliers

Code Plot

```
ggplot(office_ratings, aes(x = total_votes, y = imdb_rating)) +
  geom_jitter() +
  gghighlight(total_votes > 4000, label_key = title) +
  labs(
    title = "The Office ratings",
    x = "Total votes",
    y = "IMDB Rating"
  )
```

If you like the Dinner Party episode, I highly recommend this "oral history" of the episode published on Rolling Stone magazine.

IMDB ratings vs. seasons

Code Plot

```
ggplot(office_ratings, aes(x = factor(season), y = imdb_rating, color = season)) +
  geom_boxplot() +
  geom_jitter() +
  guides(color = FALSE) +
  labs(
    title = "The Office ratings",
    x = "Season",
    y = "IMDB Rating"
  )
```

Build recipe

Code Output

```
office_rec <- recipe(imdb_rating ~ ., data = office_train) %>%
  # title isn't a predictor, but keep around to ID
  update_role(title, new_role = "ID") %>%
  # extract month of air_date
  step_date(air_date, features = "month") %>%
  step_rm(air_date) %>%
  # make dummy variables of month
  step_dummy(contains("month")) %>%
  # remove zero variance predictors
  step_zv(all_predictors())
```

Build workflow

Code Output

```
office_wflow <- workflow() %>%
  add_model(office_mod) %>%
  add_recipe(office_rec)
```

Fit model

Code Output

```
office_fit <- office_wflow %>%
  fit(data = office_train)
```