phase1_data_loading.R

abdulbasir

2025-12-09

```
# phase 1: data loading, cleaning, and initial validation
#
# loading FDA CDER dataset, validating structure, cleaning data,
# creating response and blocking variables

# sourcing configuration and utilities
source("config.R")
source("utils.R")

# loading required libraries
suppressPackageStartupMessages({
  library(dplyr)
  library(readr)
  library(lubridate)
})

print_section_header("Phase 1: Data Loading, Cleaning, and Initial Validation")
```

```
##
## ========================================================================
## Phase 1: Data Loading, Cleaning, and Initial Validation
## ========================================================================
```

```
# 1. loading FDA CDER dataset
fda_data = load_csv(RAW_DATA_FILE)

cat(paste("Loaded", nrow(fda_data), "records from FDA CDER dataset\n"))
```

```
## Loaded 1341 records from FDA CDER dataset
```

```
cat("Columns:", paste(names(fda_data), collapse = ", "), "\n")
```

```
## Columns: Proprietary  Name, Active Ingredient/Moiety, Applicant, NDA/BLA, Application Number(1), Appl
```

```
# 2. validating data structure
missing_cols = setdiff(REQUIRED_COLUMNS, names(fda_data))
if (length(missing_cols) > 0) {
  stop(paste("Missing required columns:", paste(missing_cols, collapse = ", ")))
}

cat("All required columns present\n")
```

```
## All required columns present
```

```
cat("Date column types:\n  FDA Receipt Date:", class(fda_data$`FDA Receipt Date`),
    "\n  FDA Approval Date:", class(fda_data$`FDA Approval Date`), "\n")
```

```
## Date column types:
##   FDA Receipt Date: character
##   FDA Approval Date: character
```

```
# 3. parsing date columns
fda_data = fda_data %>%
  mutate(
    receipt_date = mdy(`FDA Receipt Date`),
    approval_date = mdy(`FDA Approval Date`)
  )

invalid_receipt = sum(is.na(fda_data$receipt_date))
invalid_approval = sum(is.na(fda_data$approval_date))

cat("Date parsing complete\nInvalid receipt dates:", invalid_receipt,
    "\nInvalid approval dates:", invalid_approval, "\n")
```

```
## Date parsing complete
## Invalid receipt dates: 0
## Invalid approval dates: 0
```

```
# 4. filtering implausible values
initial_n = nrow(fda_data)

fda_data = fda_data %>%
  mutate(review_time_days = as.numeric(approval_date - receipt_date))

implausible_negative = sum(fda_data$review_time_days < MIN_REVIEW_TIME_DAYS, na.rm = TRUE)
implausible_long = sum(fda_data$review_time_days > MAX_REVIEW_TIME_DAYS, na.rm = TRUE)

fda_data = fda_data %>%
  filter(
    review_time_days >= MIN_REVIEW_TIME_DAYS,
    review_time_days <= MAX_REVIEW_TIME_DAYS
  )

final_n = nrow(fda_data)

cat("Initial records:", initial_n, "\nImplausible records removed:", initial_n - final_n,
    "\n  - Negative review times:", implausible_negative,
    "\n  - Review times >3650 days:", implausible_long,
    "\nFinal clean records:", final_n, "\n")
```

```
## Initial records: 1341
## Implausible records removed: 6
##   - Negative review times: 0
##   - Review times >3650 days: 6
## Final clean records: 1335
```

```r
# 5. creating response variables
fda_data = fda_data %>%
  mutate(log_review_time_days = log(review_time_days))

cat(paste("Created review_time_days:", sum(!is.na(fda_data$review_time_days)), "observations\n"))
```

```
## Created review_time_days: 1335 observations
```

```r
cat(paste("Range:", round(min(fda_data$review_time_days, na.rm = TRUE)),
          "to", round(max(fda_data$review_time_days, na.rm = TRUE)), "days\n"))
```

```
## Range: 18 to 3497 days
```

```r
cat(paste("Mean:", round(mean(fda_data$review_time_days, na.rm = TRUE), 1), "days\n"))
```

```
## Mean: 545.9 days
```

```r
cat("Created log_review_time_days for scale-dependency testing\n")
```

```
## Created log_review_time_days for scale-dependency testing
```

```r
# 6. creating blocking variable (regulatory era)
fda_data = fda_data %>%
  mutate(regulatory_era = assign_regulatory_era(`Approval Year`))

cat(paste("Created regulatory_era blocking variable:", sum(!is.na(fda_data$regulatory_era)), "observati
```

```
## Created regulatory_era blocking variable: 1335 observations
```

```r
cat("Era distribution:\n")
```

```
## Era distribution:
```

```r
print(table(fda_data$regulatory_era))
```

```
##
## Early-PDUFA   Mid-PDUFA Post-FDASIA   Pre-PDUFA
##         336         262         530         207
```

```r
# 7. generating descriptive statistics
cat(paste("\nTotal observations:", nrow(fda_data), "\n"))
```

```
##
## Total observations: 1335
```

```r
cat("Date ranges:\n")
```

```
## Date ranges:
```

```r
cat(paste("  Receipt:", min(fda_data$receipt_date, na.rm = TRUE), "to",
          max(fda_data$receipt_date, na.rm = TRUE), "\n"))
```

```
##   Receipt: 1978-10-23 to 2024-05-02
```

```r
cat(paste("  Approval:", min(fda_data$approval_date, na.rm = TRUE), "to",
          max(fda_data$approval_date, na.rm = TRUE), "\n"))
```

```
##   Approval: 1985-04-09 to 2024-12-20
```

```r
cat(paste("  Years:", min(fda_data$`Approval Year`, na.rm = TRUE), "-",
          max(fda_data$`Approval Year`, na.rm = TRUE), "\n"))
```

```
##   Years: 1985 - 2024
```

```r
# 8. summary statistics for created variables
cat("\nSummary statistics for all created variables\n")
```

```
##
## Summary statistics for all created variables
```

```r
cat("review_time_days (response variable):\n")
```

```
## review_time_days (response variable):
```

```r
review_summary = summary(fda_data$review_time_days)
for (i in 1:length(review_summary)) {
  cat(sprintf("%10s: %10.2f\n", names(review_summary)[i], review_summary[i]))
}
```

```
##       Min.:      18.00
##    1st Qu.:     243.00
##     Median:     364.00
##       Mean:     545.95
##    3rd Qu.:     699.50
##       Max.:    3497.00
```

```r
cat("log_review_time_days (log-transformed response):\n")
```

```
## log_review_time_days (log-transformed response):
```

```r
log_summary = summary(fda_data$log_review_time_days)
for (i in 1:length(log_summary)) {
  cat(sprintf("%10s: %10.2f\n", names(log_summary)[i], log_summary[i]))
}
```

```
##       Min.:       2.89
##    1st Qu.:       5.49
##     Median:       5.90
##       Mean:       6.03
##    3rd Qu.:       6.55
##       Max.:       8.16
```

```r
cat("regulatory_era (blocking variable):\n")
```

```
## regulatory_era (blocking variable):
```

```r
era_counts = table(fda_data$regulatory_era)
for (era in names(era_counts)) {
  cat(sprintf("  %15s: %4d observations\n", era, era_counts[era]))
}
```

```
##       Early-PDUFA:  336 observations
##         Mid-PDUFA:  262 observations
##       Post-FDASIA:  530 observations
##         Pre-PDUFA:  207 observations
```

```r
# 9. validating temporal trends
era_means = fda_data %>%
  group_by(regulatory_era) %>%
  summarise(mean_review_time = mean(review_time_days, na.rm = TRUE), .groups = "drop")

cat("\nMean review time by regulatory era:\n")
```

```
##
## Mean review time by regulatory era:
```

```r
print(era_means)
```

```
## # A tibble: 4 x 2
##   regulatory_era mean_review_time
##   <chr>                     <dbl>
## 1 Early-PDUFA                600.
## 2 Mid-PDUFA                  469.
## 3 Post-FDASIA                384.
## 4 Pre-PDUFA                  970.
```

```r
cat("Validation:\n")
```

```
## Validation:
```

```
cat(paste("Pre-PDUFA:", round(era_means$mean_review_time[era_means$regulatory_era == "Pre-PDUFA"], 1),
          "days (expected ~970 days)\n"))
```

```
## Pre-PDUFA: 969.8 days (expected ~970 days)
```

```
cat(paste("Post-FDASIA:", round(era_means$mean_review_time[era_means$regulatory_era == "Post-FDASIA"], 1
          "days (expected <400 days)\n"))
```

```
## Post-FDASIA: 383.9 days (expected <400 days)
```

```
temporal_valid = era_means$mean_review_time[era_means$regulatory_era == "Pre-PDUFA"] >
                 era_means$mean_review_time[era_means$regulatory_era == "Post-FDASIA"]
cat(paste("Temporal trend confirmed:", temporal_valid, "\n"))
```

```
## Temporal trend confirmed: TRUE
```

```
# 10. FDA CDER dataset validation report
cat("\nDataset summary\n")
```

```
##
## Dataset summary
```

```
cat(paste("   Total records after cleaning:", final_n, "\n"))
```

```
##    Total records after cleaning: 1335
```

```
cat(paste("   Original records:", initial_n, "\n"))
```

```
##    Original records: 1341
```

```
cat(paste("   Records removed:", initial_n - final_n, "(implausible review times)\n"))
```

```
##    Records removed: 6 (implausible review times)
```

```
cat("Date ranges\n")
```

```
## Date ranges
```

```
cat(paste("   FDA Receipt Date:", min(fda_data$receipt_date, na.rm = TRUE), "to",
          max(fda_data$receipt_date, na.rm = TRUE), "\n"))
```

```
##    FDA Receipt Date: 1978-10-23 to 2024-05-02
```

```
cat(paste("   FDA Approval Date:", min(fda_data$approval_date, na.rm = TRUE), "to",
          max(fda_data$approval_date, na.rm = TRUE), "\n"))
```

```
##    FDA Approval Date: 1985-04-09 to 2024-12-20
```

```
cat(paste("   Approval Years:", min(fda_data$`Approval Year`, na.rm = TRUE), "-",
          max(fda_data$`Approval Year`, na.rm = TRUE), "\n"))
```

```
##    Approval Years: 1985 - 2024
```

```
cat("Data completeness rates (% non-null)\n")
```

```
## Data completeness rates (% non-null)
```

```
for (col in REQUIRED_COLUMNS) {
  completeness = (1 - sum(is.na(fda_data[[col]])) / nrow(fda_data)) * 100
  cat(sprintf("   %s: %.1f%%\n", col, completeness))
}
```

```
##    FDA Receipt Date: 100.0%
##    FDA Approval Date: 100.0%
##    Review Designation: 100.0%
##    Abbreviated Indication(s): 49.4%
##    Orphan Drug Designation: 100.0%
##    Accelerated Approval: 100.0%
##    Breakthrough Therapy Designation: 100.0%
##    Fast Track Designation: 100.0%
##    Qualified Infectious Disease Product: 100.0%
##    Approval Year: 100.0%
```

```
cat("Data quality issues identified\n")
```

```
## Data quality issues identified
```

```
cat(paste("   •", initial_n - final_n, "records removed with review times >3650 days (biologically impla
```

```
##    • 6 records removed with review times >3650 days (biologically implausible)
```

```
cat("   • All date fields successfully parsed\n")
```

```
##    • All date fields successfully parsed
```

```
cat("Validation complete - dataset ready for analysis\n")
```

```
## Validation complete - dataset ready for analysis
```

```
save_csv(fda_data, CLEAN_DATA_FILE)
```

```
## saving results to: /Users/abdulbasir/Downloads/Experimental AI/fda-oncology-approval-analysis/results
```

```r
cat("\nPhase 1 complete\n")
```

```
##
## Phase 1 complete
```