PID-5 EMA — Build analysis dataset (baseline + EMA + exam tags)

Clean, tag, QA, and export

Corrado Caudek

Questa notebook costruisce **unico file analitico** per il progetto *PID-5 EMA*, unendo:
1) **baseline** (PID-5 domini senza item EMA, ESI-BF, DASS-21 baseline),
2) **EMA** (misure dinamiche, incluse `happy/sad/satisfied/angry`, SCS di stato),
3) **tag esame** (`exam_period`: baseline / pre_exam / post_exam),
4) **filtri qualità** riproducibili.
Alla fine, salva `{params.export_path}` e stampa **log di inclusioni/esclusioni**.

1 Setup

```
suppressPackageStartupMessages({
  library(tidyverse)
  library(here)
  library(rio)
  library(stringr)
  library(lubridate)
  library(forcats)
  library(broom)
  library(glue)
  library(conflicted)
})

conflict_prefer("filter", "dplyr")
```

```
[conflicted] Will prefer dplyr::filter over any other package.
```

```
conflict_prefer("select", "dplyr")
```

```
[conflicted] Will prefer dplyr::select over any other package.
```

```
conflict_prefer("sd", "stats")
```

```
[conflicted] Will prefer stats::sd over any other package.
```

```
conflict_prefer("var", "stats")
```

```
[conflicted] Will prefer stats::var over any other package.
```

```
theme_set(theme_minimal(base_size = 12))
options(dplyr.summarise.inform = FALSE)
```

1.1 Utility (coercioni sicure, somma NA-robusta, clamp)

```
# Coercizione numerica "safe"
to_num <- function(x) suppressWarnings(as.numeric(x))

# Clamp in [0, 100]
clamp_0_100 <- function(x) {
  x <- to_num(x)
  ifelse(is.na(x), NA_real_, pmin(pmax(x, 0), 100))
}

# Somma riga con NA-robustezza ma torna NA se tutti gli addendi sono NA
row_sum_na_all_missing_na <- function(mat) {
  s <- rowSums(mat, na.rm = TRUE)
  all_na <- rowSums(!is.na(mat)) == 0
  s[all_na] <- NA_real_
  s
}
```

2 1) Import baseline: ESI-BF e PID-5 (domini, senza item EMA)

```
# ESI-BF (già ricodificato e con user_id pulito dallo script `import_esi.R`)
esi_bf <- rio::import(here::here("data","processed","esi_bf.csv")) %>%
  distinct(user_id, .keep_all = TRUE) %>%
  select(user_id, esi_bf)
stopifnot(nrow(esi_bf) > 0, all(c("user_id","esi_bf") %in% names(esi_bf)))

# PID-5 (domini senza item EMA) da `import_pid5_subset.R`
pid5 <- rio::import(here::here("data","processed","pid5.csv")) %>%
  distinct(user_id, .keep_all = TRUE) %>%
  select(user_id, starts_with("domain_"))
stopifnot(nrow(pid5) > 0)

# Merge baseline
baseline <- left_join(esi_bf, pid5, by = "user_id")
```

2.1 1.1) Escludi careless responding (lista definita a priori)

```
user_id_with_careless_responding <- c(
  "ma_se_2005_11_14_490","reve20041021036","di_ma_2005_10_20_756","pa_sc_2005_09_10_468",
  "il_re_2006_01_18_645","so_ma_2003_10_13_804","lo_ca_2005_05_07_05_437",
  "va_ma_2005_05_31_567","no_un_2005_06_29_880","an_bo_1988_08_24_166",
  "st_ma_2004_04_21_426","an_st_2005_10_16_052","vi_de_2002_12_30_067",
  "gi_ru_2005_03_08_033","al_mi_2005_03_05_844","la_ma_2006_01_31_787",
  "gi_lo_2004_06_27_237","ch_bi_2001_01_28_407","al_pe_2001_04_20_079",
  "le_de_2003_09_05_067","fe_gr_2002_02_19_434","ma_ba_2002_09_09_052",
```

```
  "ca_gi_2003_09_16_737","an_to_2003_08_06_114","al_se_2003_07_28_277",
  "ja_tr_2002_10_06_487","el_ci_2002_02_15_057","se_ti_2000_03_04_975",
  "co_ga_2003_10_29_614","al_ba_2003_18_07_905","bi_ro_2003_09_07_934",
  "an_va_2004_04_08_527","ev_cr_2003_01_27_573"
)

baseline <- baseline %>%
  mutate(flag_careless = user_id %in% user_id_with_careless_responding)

baseline_keep <- baseline %>% filter(!flag_careless)
baseline_drop <- baseline %>% filter(flag_careless) %>% select(user_id)
```

3 2) Import EMA e merge con baseline

```
ema_raw <- readRDS(here::here("data","raw","ema","ema_data_scoring.RDS")) %>%
  rename(user_id = subj_code)

df0 <- left_join(baseline_keep %>% select(-flag_careless), ema_raw, by="user_id")
n_distinct(df0$user_id) -> n_users_after_careless
glue("Soggetti dopo esclusione careless: {n_users_after_careless}") %>% message()
```

```
Soggetti dopo esclusione careless: 429
```

3.1 2.1) Filtro per numerosità EMA (riportabile)

```
counts <- df0 %>% count(user_id, name = "n_ema")
keep_ids <- counts %>% filter(n_ema >= 5 & n_ema <= 40) %>% pull(user_id)

df1 <- df0 %>% filter(user_id %in% keep_ids)

log_n <- tibble(
  step = c("after_careless","after_min_n_ema"),
  n_subjects = c(n_users_after_careless, n_distinct(df1$user_id))
)
log_n
```

```
# A tibble: 2 x 2
  step            n_subjects
  <chr>                <int>
1 after_careless         429
2 after_min_n_ema        379
```

4 3) Import metadati corso/sesso e tag esami

```
exam_data <- rio::import(here::here("data","raw","all_combined_sex_NEW_1.xlsx")) %>%
  select(user_id = subj_code, course, sex)
```

```r
df2 <- left_join(df1, exam_data, by="user_id")
```

4.1 3.1) Crea exam_period da calendario corsi

```r
# Parametri dalle YAML params
psico_pre   <- as.Date(params$psico_pre)
psico_post  <- as.Date(params$psico_post)
test_pre    <- as.Date(params$test_pre)
test_post   <- as.Date(params$test_post)
interv_pre  <- as.Date(params$interv_pre)
interv_post <- as.Date(params$interv_post)

df2 <- df2 %>%
  mutate(
    day = as.Date(day),
    exam_period = case_when(
      course == "Clinica" ~ "baseline",
      course == "Psicometria" & day %in% psico_pre  ~ "pre_exam",
      course == "Psicometria" & day %in% psico_post ~ "post_exam",
      course == "Testing"     & day %in% test_pre   ~ "pre_exam",
      course == "Testing"     & day %in% test_post  ~ "post_exam",
      course == "Interventi"  & day %in% interv_pre ~ "pre_exam",
      course == "Interventi"  & day %in% interv_post~ "post_exam",
      TRUE ~ "baseline"
    ),
    exam_period = factor(exam_period, levels = c("baseline","pre_exam","post_exam"))
  )

count(df2, course, exam_period)
```

```
        course exam_period    n
1      Clinica    baseline 3853
2   Interventi    baseline  394
3   Interventi    pre_exam   18
4   Interventi   post_exam   17
5  Psicometria    baseline 3475
6  Psicometria    pre_exam  272
7  Psicometria   post_exam  228
8      Testing    baseline 1344
9      Testing    pre_exam  117
10     Testing   post_exam  111
11        <NA>    baseline  144
```

5 4) Costruisci SCS di stato (cs_pos, ucs_neg) dalle colonne scs*_pos/neg

```r
pos_items <- names(df2)[str_detect(names(df2), "^scs\\d+_pos$")]
neg_items <- names(df2)[str_detect(names(df2), "^scs\\d+_neg$")]
stopifnot(length(pos_items) > 0, length(neg_items) > 0)

df2[pos_items] <- lapply(df2[pos_items], to_num)
df2[neg_items] <- lapply(df2[neg_items], to_num)

df2 <- df2 %>% mutate(
  cs_pos  = row_sum_na_all_missing_na(as.matrix(pick(all_of(pos_items)))),
  ucs_neg = row_sum_na_all_missing_na(as.matrix(pick(all_of(neg_items))))
)
```

6 5) Negative affect EMA da happy/satisfied/sad/angry (reverse su positivi)

```r
df2 <- df2 %>%
  mutate(
    happy     = clamp_0_100(happy),
    satisfied = clamp_0_100(satisfied),
    sad       = clamp_0_100(sad),
    angry     = clamp_0_100(angry),
    happy_rc     = if_else(is.na(happy),     NA_real_, 100 - happy),
    satisfied_rc = if_else(is.na(satisfied), NA_real_, 100 - satisfied)
  ) %>%
  mutate(
    negative_affect_ema_raw = row_sum_na_all_missing_na(as.matrix(pick(happy_rc, satisfied_r
  ) %>%
  mutate(
    neg_affect_ema = as.numeric(scale(negative_affect_ema_raw))
  ) %>%
  select(-happy_rc, -satisfied_rc)
```

7 6) QA essenziale (riportabile nel manoscritto)

```r
qa_subjects <- df2 %>%
  group_by(user_id) %>%
  summarise(
    n_ema = n(),
    n_days = n_distinct(day),
    any_pre  = any(exam_period == "pre_exam", na.rm = TRUE),
    any_post = any(exam_period == "post_exam", na.rm = TRUE),
    any_base = any(exam_period == "baseline", na.rm = TRUE)
  ) %>%
  arrange(desc(n_ema))

print(qa_subjects, n = 10)
```

```
# A tibble: 379 x 6
   user_id              n_ema n_days any_pre any_post any_base
   <chr>                <int>  <int> <lgl>   <lgl>    <lgl>
 1 el_to_2005_09_02_232    37     31 TRUE    TRUE     TRUE
 2 ad_pa_2006_01_27_630    31     30 TRUE    TRUE     TRUE
 3 al_lo_1999_09_19_127    31     31 TRUE    TRUE     TRUE
 4 al_mo_2005_10_03_273    31     30 TRUE    TRUE     TRUE
 5 al_za_2005_07_11_637    31     29 TRUE    TRUE     TRUE
 6 an_ba_2001_12_27_169    31     30 TRUE    TRUE     TRUE
 7 an_ma_1995_03_09_029    31     31 TRUE    TRUE     TRUE
 8 ca_ro_2004_09_03_927    31     30 TRUE    TRUE     TRUE
 9 ca_te_2005_07_04_549    31     30 TRUE    TRUE     TRUE
10 ca_za_2005_06_29_989    31     31 TRUE    TRUE     TRUE
# i 369 more rows
```

```r
qa_missing <- df2 %>%
  summarise(
    pct_missing_neg_affect = mean(is.na(neg_affect_ema))*100,
    pct_missing_cs_pos     = mean(is.na(cs_pos))*100,
    pct_missing_ucs_neg    = mean(is.na(ucs_neg))*100
  )
qa_missing
```

```
  pct_missing_neg_affect pct_missing_cs_pos pct_missing_ucs_neg
1             0.02005415           37.87226            37.87226
```

7.1 6.1) Effetti del periodo d'esame stratificati per sesso (descrittivi)

```r
desc_by_sex <- df2 %>%
  filter(exam_period %in% c("pre_exam","post_exam")) %>%
  group_by(sex, exam_period) %>%
  summarise(
    n = n(),
    neg_aff_mean = mean(neg_affect_ema, na.rm = TRUE),
    cs_pos_mean  = mean(cs_pos, na.rm = TRUE),
    ucs_neg_mean = mean(ucs_neg, na.rm = TRUE)
  ) %>%
  arrange(sex, exam_period)

desc_by_sex
```

```
# A tibble: 4 x 6
# Groups:   sex [2]
  sex     exam_period     n neg_aff_mean cs_pos_mean ucs_neg_mean
  <chr>   <fct>       <int>        <dbl>       <dbl>        <dbl>
1 Femmina pre_exam      340        0.421        1.35         0.182
2 Femmina post_exam     295       -0.257        2.45        -1.78
3 Maschio pre_exam       67        0.231       0.646        -1.78
4 Maschio post_exam      61       -0.0286       1.86        -3.14
```

8 7) Esporta dataset finale + log esclusioni

```
# Log esclusioni
log_exclusions <- list(
  careless_ids = baseline_drop,
  min_n_ema_threshold = tibble(min_n_ema = params$min_n_ema,
                               dropped_ids = setdiff(unique(df0$user_id), unique(df1$user_id
)

# Dataset finale con rinomina PID-5 baseline
dat_final <- df2 %>%
  rename(
    esi_bf_baseline = esi_bf,
    pid5_negative_affect_baseline = domain_negative_affect,
    pid5_detachment_baseline      = domain_detachment,
    pid5_antagonism_baseline      = domain_antagonism,
    pid5_disinhibition_baseline   = domain_disinhibition,
    pid5_psychoticism_baseline    = domain_psychoticism
  )

# Salva
rio::export(dat_final, here::here(params$export_path))

cat(glue("\nFile esportato: {here::here(params$export_path)}\n"))
```

```
File esportato: /Users/corrado/_repositories/pid5-ema/data/processed/ema_plus_baseline_exam
```

```
# Mostra dimensioni e anteprima
dim(dat_final)
```

```
[1] 9973   94
```

```
dat_final %>% arrange(user_id, day, hour) %>% slice_head(n = 8)
```

```
              user_id esi_bf_baseline pid5_negative_affect_baseline
1 ad_pa_2006_01_27_630              12                            44
2 ad_pa_2006_01_27_630              12                            44
3 ad_pa_2006_01_27_630              12                            44
4 ad_pa_2006_01_27_630              12                            44
5 ad_pa_2006_01_27_630              12                            44
6 ad_pa_2006_01_27_630              12                            44
7 ad_pa_2006_01_27_630              12                            44
8 ad_pa_2006_01_27_630              12                            44
  pid5_detachment_baseline pid5_antagonism_baseline pid5_disinhibition_baseline
1                        7                       13                           32
2                        7                       13                           32
3                        7                       13                           32
4                        7                       13                           32
5                        7                       13                           32
6                        7                       13                           32
7                        7                       13                           32
```

```
8                        7                       13                          32
  pid5_psychoticism_baseline happy sad satisfied angry pid5_13 pid5_15 pid5_11
1                         43   100   0        64    16       0       0       0
2                         43    85  11        82     0       0       1       1
3                         43    89  15        32   100       0       0       2
4                         43    71  16        54     0       0       0       1
5                         43    93  17        40    67       0       0       2
6                         43    88   0        72    61       0       0       2
7                         43    83   0        73    50       0       0       3
8                         43   100   0        87    29       0       0       2
  pid5_3 pid5_2 pid5_7 pid5_14 pid5_6 pid5_4 pid5_12 pid5_1 pid5_9 pid5_5
1      0      0      1       0      0      0       0      0      0      0
2      2      2      0       0      0      0       0      0      0      0
3      3      3      0       0      0      0       0      0      0      0
4      2      3      0       0      0      0       0      0      0      0
5      3      3      0       0      0      0       1      1      0      0
6      2      2      0       0      0      0       0      0      0      0
7      3      2      0       0      0      0       0      0      0      0
8      3      2      0       0      0      0       0      1      0      0
  pid5_8 pid5_10 tripm_1 tripm_3 tripm_2 tripm_4 dass21_2 dass21_5 dass21_3
1      0       2       3       1       2       3        0        0        0
2      0       2       3       1       1       4       NA       NA       NA
3      0       2       3       1       1       4        1        0        0
4      0       0       2       1       1       3       NA       NA       NA
5      0       1       1       1       1       4       NA       NA       NA
6      0       0       1       1       1       4       NA       NA       NA
7      0       0       3       1       1       4        1        0        1
8      0       0       3       1       1       4        1        0        1
  dass21_4 dass21_6 dass21_1 scs3_pos scs5_neg scs7_pos scs4_neg scs2_neg
1        0        0        0        1       -3        2        1       -3
2       NA       NA       NA       NA       NA       NA       NA       NA
3        0        0        2        3       -3        2        1       -2
4       NA       NA       NA       NA       NA       NA       NA       NA
5       NA       NA       NA       NA       NA       NA       NA       NA
6       NA       NA       NA       NA       NA       NA       NA       NA
7        1        1        1        2        2        1        3       -1
8        0        0        0        2       -3        2        3       -1
  scs1_pos scs8_neg scs6_pos vq_2 vq_1 vq_4 vq_3 cope_nvi_2 cope_nvi_7
1        2        2       -3   NA   NA   NA   NA         NA         NA
2       NA       NA       NA    0    0    0    0          1          3
3        2        2        1   NA   NA   NA   NA         NA         NA
4       NA       NA       NA    0    0    0    0          2          2
5       NA       NA       NA    0    0    0    0          2          1
6       NA       NA       NA    0    0    0    0          2          2
7        1        3        2   NA   NA   NA   NA         NA         NA
8        2        3       -1   NA   NA   NA   NA         NA         NA
  cope_nvi_5 cope_nvi_8 cope_nvi_9 cope_nvi_10 cope_nvi_1 cope_nvi_3 cope_nvi_4
1         NA         NA         NA          NA         NA         NA         NA
2          3          3          3           2          2          3          3
```

```
3         NA         NA         NA          NA         NA         NA         NA
4          4          3          2           2          1          2          3
5          3          3          1           2          4          2          3
6          2          2          2           3          2          2          2
7         NA         NA         NA          NA         NA         NA         NA
8         NA         NA         NA          NA         NA         NA         NA
  cope_nvi_6        day hour calendar_day bysubj_day context_quality
1         NA 2025-03-12   18            1          1               1
2          3 2025-03-15    9            2          2              -1
3         NA 2025-03-19   18            3          3              -1
4          3 2025-03-22   21            4          4               0
5          3 2025-03-26   18            5          5               0
6          4 2025-03-29    9            6          6               2
7         NA 2025-04-02   20            7          7               1
8         NA 2025-04-05   12            8          8               2
  context_control context_support context_threat pid5_sum
1               4               4              1        3
2               4               4              2        8
3               0               5              4       10
4               3               3              2        6
5               3               4              3       11
6               4               4              0        6
7               5               5              0        8
8               5               5              0        8
  pid5_negative_affectivity pid5_detachment pid5_antagonism pid5_disinhibition
1                         0               0               1                  2
2                         4               0               0                  3
3                         6               0               0                  4
4                         5               0               0                  1
5                         7               0               0                  4
6                         4               0               0                  2
7                         5               0               0                  3
8                         6               0               0                  2
  pid5_psychoticism ipv_sum tripm_4_rev tripm_sum tripm_boldness tripm_meanness
1                 0      NA           2         9              5               3
2                 1       0           1         9              4               2
3                 0      NA           1         9              4               2
4                 0       0           2         7              3               3
5                 0       0           1         7              2               2
6                 0       0           1         7              2               2
7                 0      NA           1         9              4               2
8                 0      NA           1         9              4               2
  dass_sum dass_stress dass_depression dass_anxiety cope_10_rev cope_avoid
1        0           0               0            0         NA         NA
2       NA          NA              NA           NA          3          3
3        3           3               0            0         NA         NA
4       NA          NA              NA           NA          3          3
5       NA          NA              NA           NA          3          6
6       NA          NA              NA           NA          2          4
```

```
7        5           2               2           1          NA         NA
8        2           1               1           0          NA         NA
  cope_prob_or cope_social_support cope_positive_att cope_trascendent_or cs_pos
1           NA                  NA                NA                   NA      2
2            6                   6                 6                    6     NA
3           NA                  NA                NA                   NA      8
4            5                   7                 5                    5     NA
5            5                   6                 4                    4     NA
6            4                   6                 4                    4     NA
7           NA                  NA                NA                   NA      6
8           NA                  NA                NA                   NA      5
  ucs_neg      course     sex exam_period negative_affect_ema_raw
1      -3 Psicometria Femmina    baseline                      52
2      NA Psicometria Femmina    baseline                      44
3      -2 Psicometria Femmina    baseline                     194
4      NA Psicometria Femmina    baseline                      91
5      NA Psicometria Femmina    baseline                     151
6      NA Psicometria Femmina    baseline                     101
7       7 Psicometria Femmina    baseline                      94
8       2 Psicometria Femmina    baseline                      42
  neg_affect_ema
1    -1.16610295
2    -1.26218787
3     0.53940443
4    -0.69768895
5     0.02294797
6    -0.57758279
7    -0.66165710
8    -1.28620910
```

9 8) Nota su DASS-21 baseline (opzionale)

Se desideri aggiungere i punteggi **DASS-21 baseline** calcolati dalle 21 risposte testuali, aggiungi il blocco dedicato (vedi la tua sezione *Add DASS-21*) e fai `left_join()` su `user_id` prima dell'export.