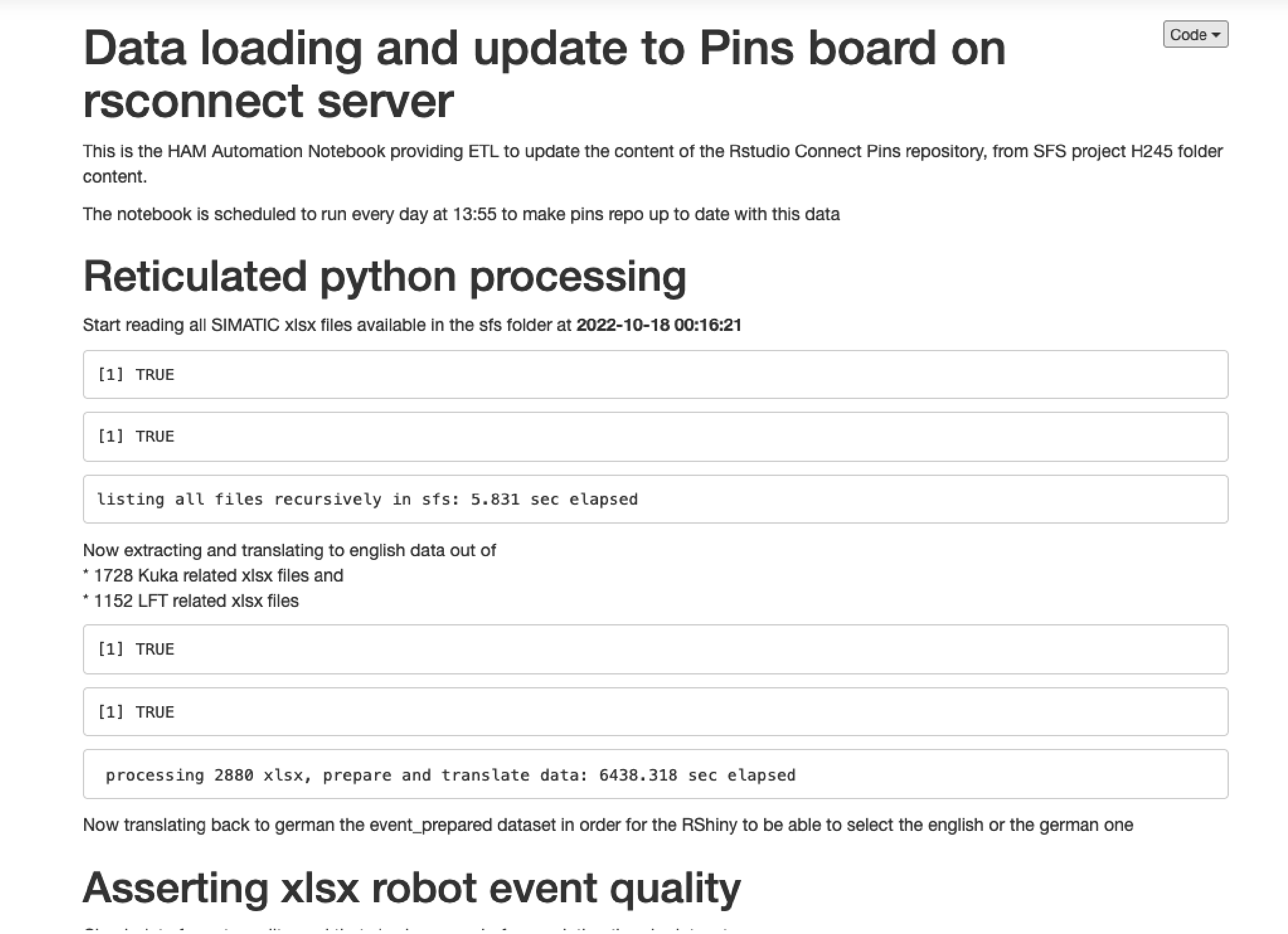
{pointblank}

How to shine with data-quality checks

Advanced Analytics & Artificial Intelligence PSL

Make your data fly with Advanced Analytics & AI

Before poinblank : a true story



Data loading and update to Pins board on rsconnect server

This is the HAM Automation Notebook providing ETL to update the content of the Rstudio Connect Pins repository, from SFS project H245 folder content.

The notebook is scheduled to run every day at 13:55 to make pins repo up to date with this data

Reticulated python processing

Start reading all SIMATIC xlsx files available in the sfs folder at **2022-10-18 00:16:21**

```
[1] TRUE
```

```
[1] TRUE
```

```
listing all files recursively in sfs: 5.831 sec elapsed
```

Now extracting and translating to english data out of
* 1728 Kuka related xlsx files and
* 1152 LFT related xlsx files

```
[1] TRUE
```

```
[1] TRUE
```

```
 processing 2880 xlsx, prepare and translate data: 6438.318 sec elapsed
```

Now translating back to german the event_prepared dataset in order for the RShiny to be able to select the english or the german one

Asserting xlsx robot event quality

Enter a MSN: 10391

Production Line: H245 - MCA

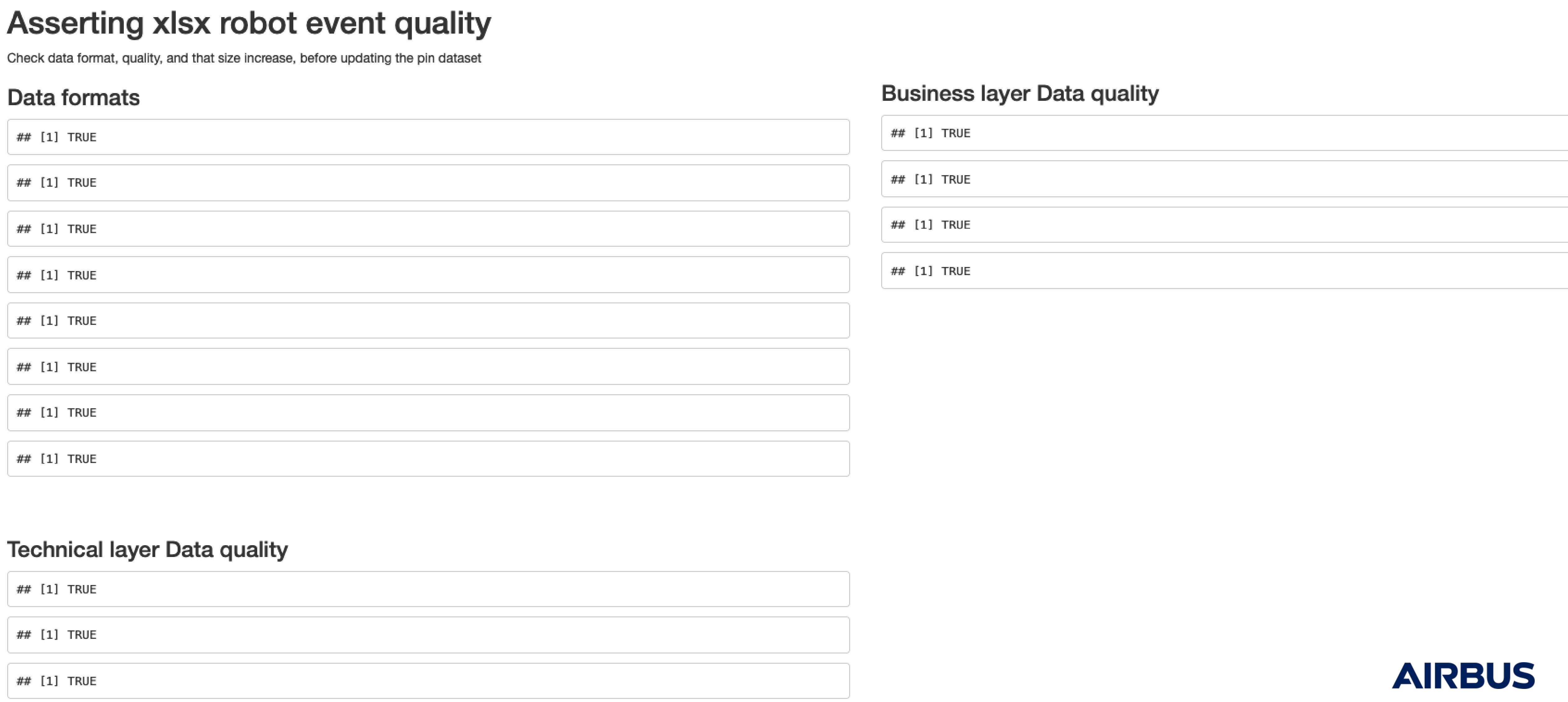
Station: ST74G, ST74F, ST74H, All stations

Localization: Lower-Left, Upper-Right, Upper-Left, Lower-Right, All

Range of dates: 2020-07-11 - 2020-10-0

Type of visualization: ● Waterfall ○ Pareto

Before poinblank



Asserting xlsx robot event quality

Check data format, quality, and that size increase, before updating the pin dataset

Data formats

```
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] TRUE
```

Technical layer Data quality

```
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] TRUE
```

Business layer Data quality

```
## [1] TRUE
```

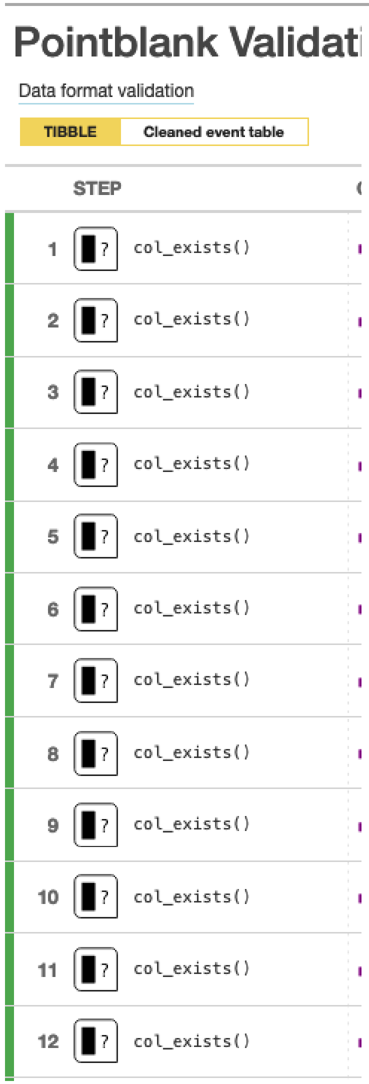
```
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] TRUE
```

After poinblank

Data formats

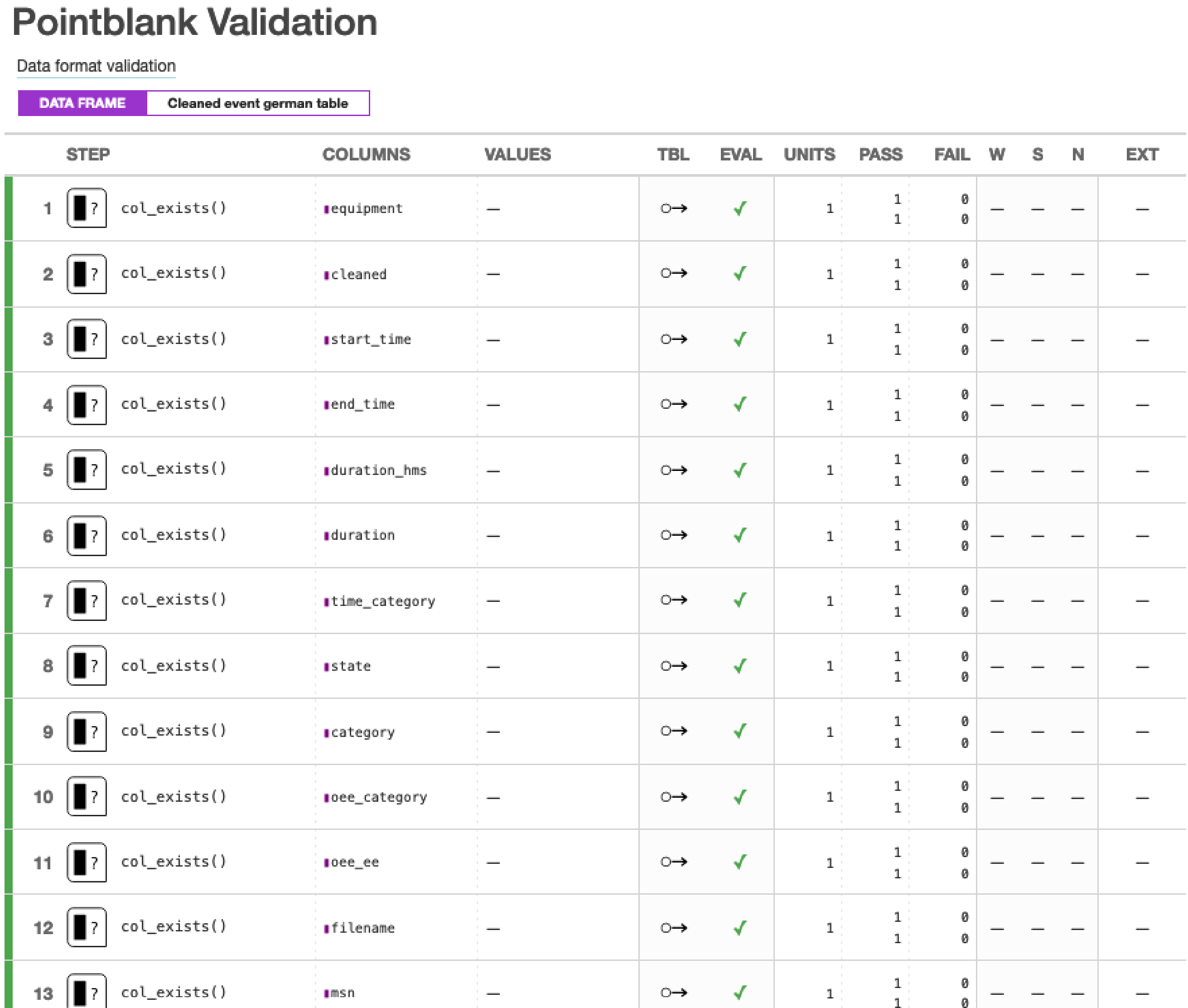


Pointblank Validati

Data format validation

TIBBLE | Cleaned event table

	STEP
1	col_exists()
2	col_exists()
3	col_exists()
4	col_exists()
5	col_exists()
6	col_exists()
7	col_exists()
8	col_exists()
9	col_exists()
10	col_exists()
11	col_exists()
12	col_exists()



Pointblank Validation

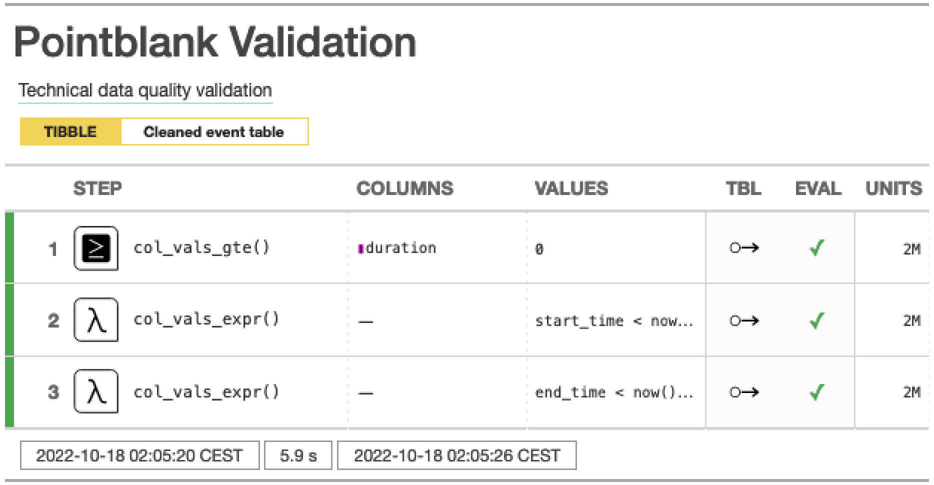
Data format validation

DATA FRAME | Cleaned event german table

	STEP	COLUMNS	VALUES	TBL	EVAL	UNITS	PASS	FAIL	W	S	N	EXT
1	col_exists()	equipment	—	○→	✓	1	1 1	0 0	—	—	—	—
2	col_exists()	cleaned	—	○→	✓	1	1 1	0 0	—	—	—	—
3	col_exists()	start_time	—	○→	✓	1	1 1	0 0	—	—	—	—
4	col_exists()	end_time	—	○→	✓	1	1 1	0 0	—	—	—	—
5	col_exists()	duration_hms	—	○→	✓	1	1 1	0 0	—	—	—	—
6	col_exists()	duration	—	○→	✓	1	1 1	0 0	—	—	—	—
7	col_exists()	time_category	—	○→	✓	1	1 1	0 0	—	—	—	—
8	col_exists()	state	—	○→	✓	1	1 1	0 0	—	—	—	—
9	col_exists()	category	—	○→	✓	1	1 1	0 0	—	—	—	—
10	col_exists()	oee_category	—	○→	✓	1	1 1	0 0	—	—	—	—
11	col_exists()	oee_ee	—	○→	✓	1	1 1	0 0	—	—	—	—
12	col_exists()	filename	—	○→	✓	1	1 1	0 0	—	—	—	—
13	col_exists()	msn	—	○→	✓	1	1 1	0 0	—	—	—	—

After poinblank

Technical layer Data quality



Pointblank Validation

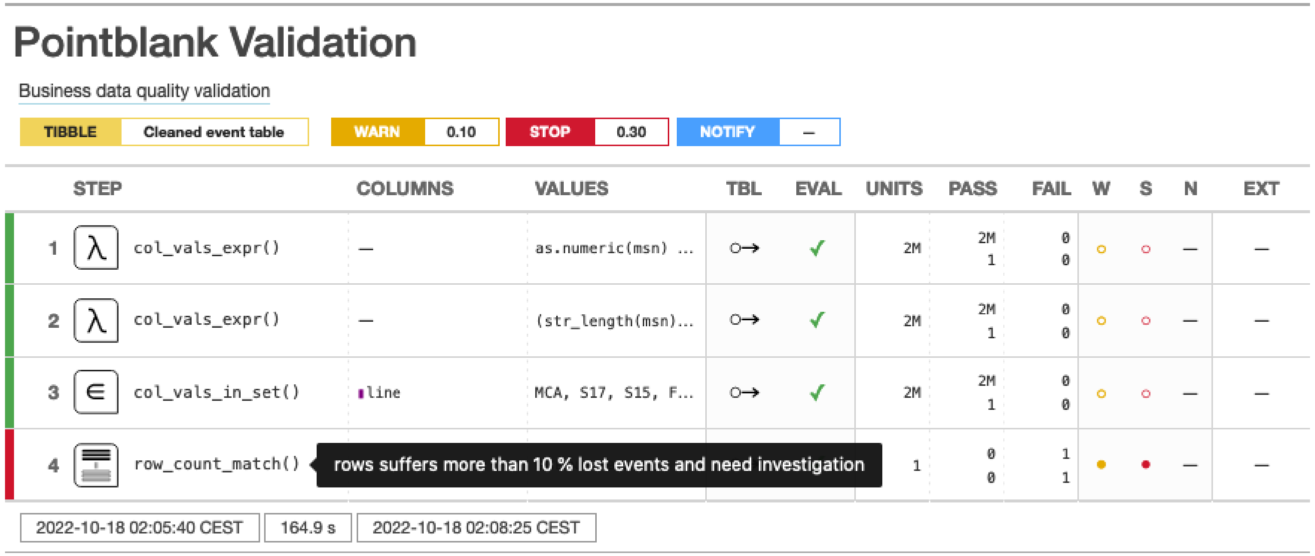
Technical data quality validation

TIBBLE | Cleaned event table

	STEP	COLUMNS	VALUES	TBL	EVAL	UNITS
1	≥ col_vals_gte()	▮duration	0	○→	✓	2M
2	λ col_vals_expr()	—	start_time < now…	○→	✓	2M
3	λ col_vals_expr()	—	end_time < now()…	○→	✓	2M

2022-10-18 02:05:20 CEST | 5.9 s | 2022-10-18 02:05:26 CEST

Business layer Data quality



Pointblank Validation

Business data quality validation

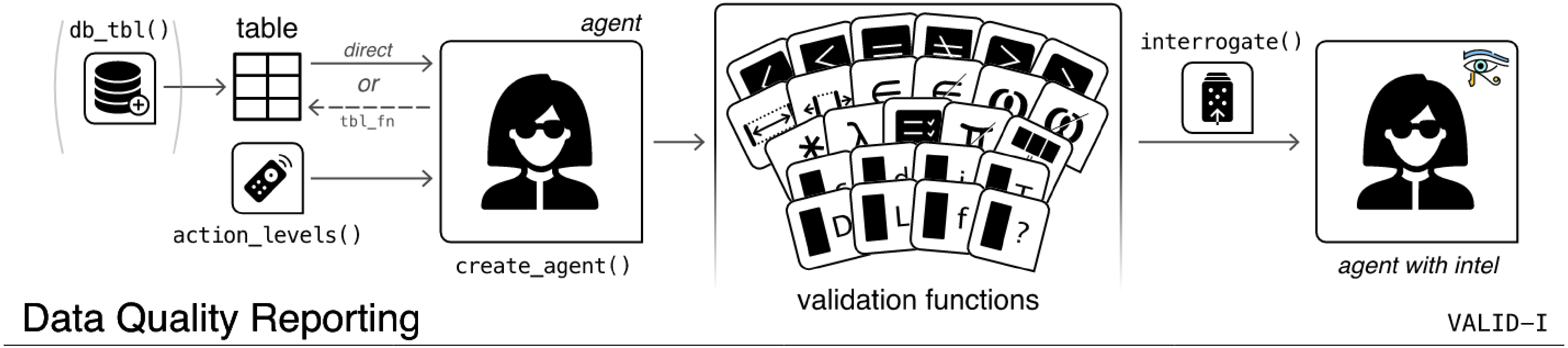
TIBBLE | Cleaned event table WARN | 0.10 STOP | 0.30 NOTIFY | —

	STEP	COLUMNS	VALUES	TBL	EVAL	UNITS	PASS	FAIL	W	S	N	EXT
1	λ col_vals_expr()	—	as.numeric(msn) …	○→	✓	2M	2M 1	0 0	○	○	—	—
2	λ col_vals_expr()	—	(str_length(msn)…	○→	✓	2M	2M 1	0 0	○	○	—	—
3	∈ col_vals_in_set()	▮line	MCA, S17, S15, F…	○→	✓	2M	2M 1	0 0	○	○	—	—
4	row_count_match()					1	0 0	1 1	●	●	—	—

rows suffers more than 10 % lost events and need investigation

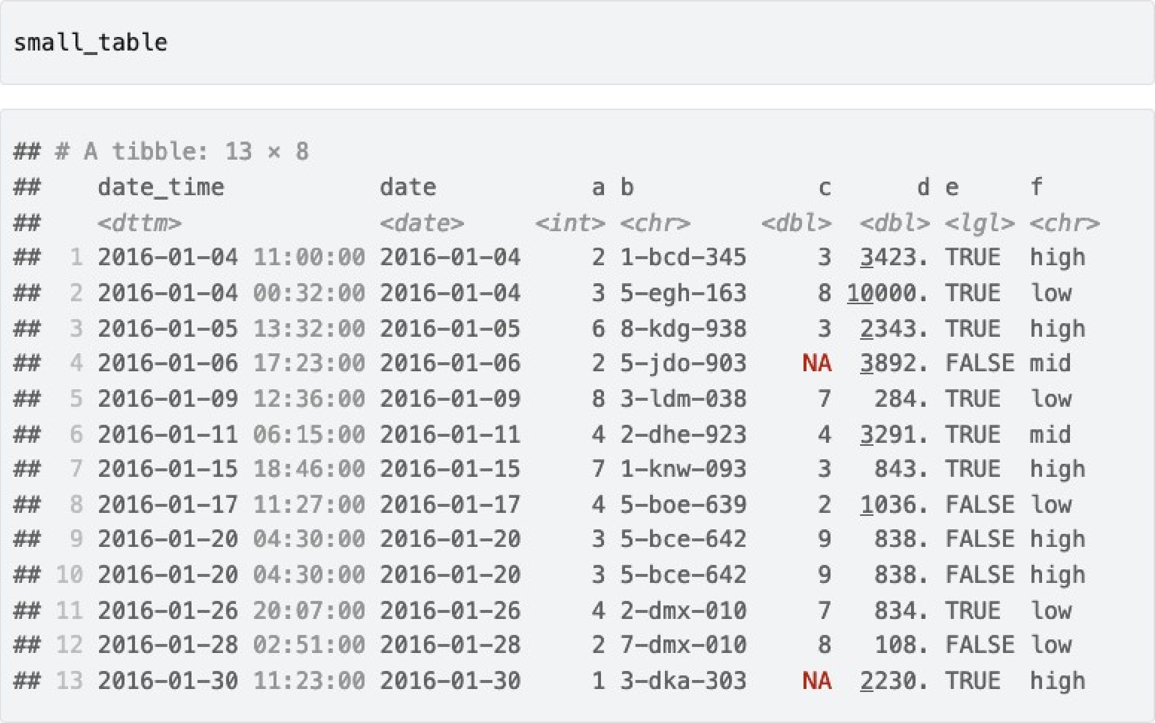
2022-10-18 02:05:40 CEST | 164.9 s | 2022-10-18 02:08:25 CEST

Pointblank : Dataset validation setup logic



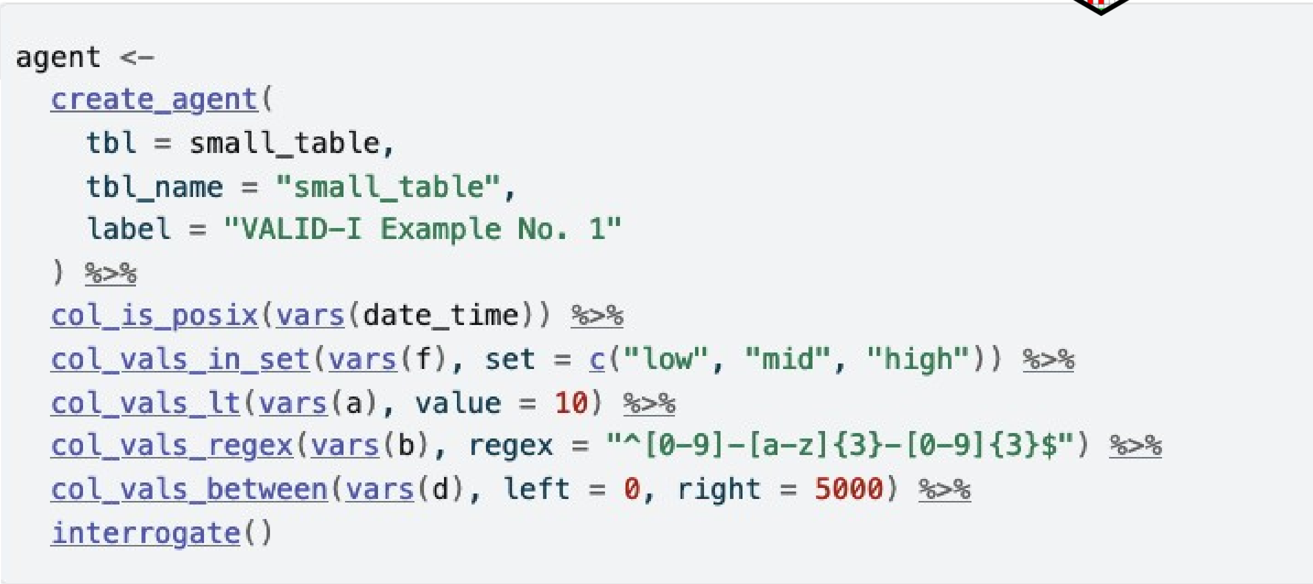
From https://rich-iannone.github.io/pointblank/articles/validation_workflows.html

Pointblank : basic example

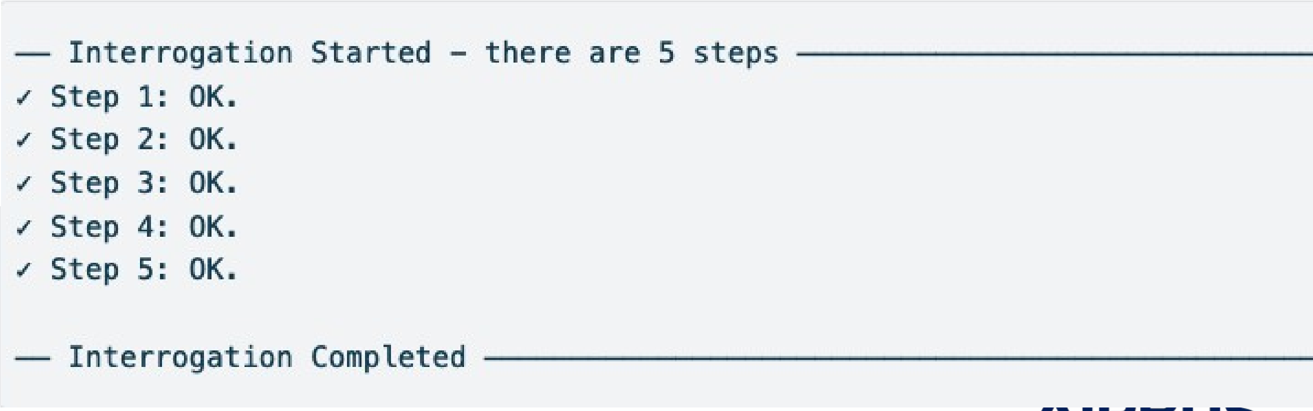


```
small_table
```

```
## # A tibble: 13 × 8
##    date_time           date           a b           c      d e     f
##    <dttm>              <date>     <int> <chr>   <dbl>  <dbl> <lgl> <chr>
##  1 2016-01-04 11:00:00 2016-01-04     2 1-bcd-345     3  3423. TRUE  high
##  2 2016-01-04 00:32:00 2016-01-04     3 5-egh-163     8 10000. TRUE  low
##  3 2016-01-05 13:32:00 2016-01-05     6 8-kdg-938     3  2343. TRUE  high
##  4 2016-01-06 17:23:00 2016-01-06     2 5-jdo-903    NA  3892. FALSE mid
##  5 2016-01-09 12:36:00 2016-01-09     8 3-ldm-038     7   284. TRUE  low
##  6 2016-01-11 06:15:00 2016-01-11     4 2-dhe-923     4  3291. TRUE  mid
##  7 2016-01-15 18:46:00 2016-01-15     7 1-knw-093     3   843. TRUE  high
##  8 2016-01-17 11:27:00 2016-01-17     4 5-boe-639     2  1036. FALSE low
##  9 2016-01-20 04:30:00 2016-01-20     3 5-bce-642     9   838. FALSE high
## 10 2016-01-20 04:30:00 2016-01-20     3 5-bce-642     9   838. FALSE high
## 11 2016-01-26 20:07:00 2016-01-26     4 2-dmx-010     7   834. TRUE  low
## 12 2016-01-28 02:51:00 2016-01-28     2 7-dmx-010     8   108. FALSE low
## 13 2016-01-30 11:23:00 2016-01-30     1 3-dka-303    NA  2230. TRUE  high
```



```
agent <-
  create_agent(
    tbl = small_table,
    tbl_name = "small_table",
    label = "VALID-I Example No. 1"
  ) %>%
  col_is_posix(vars(date_time)) %>%
  col_vals_in_set(vars(f), set = c("low", "mid", "high")) %>%
  col_vals_lt(vars(a), value = 10) %>%
  col_vals_regex(vars(b), regex = "^[0-9]-[a-z]{3}-[0-9]{3}$") %>%
  col_vals_between(vars(d), left = 0, right = 5000) %>%
  interrogate()
```

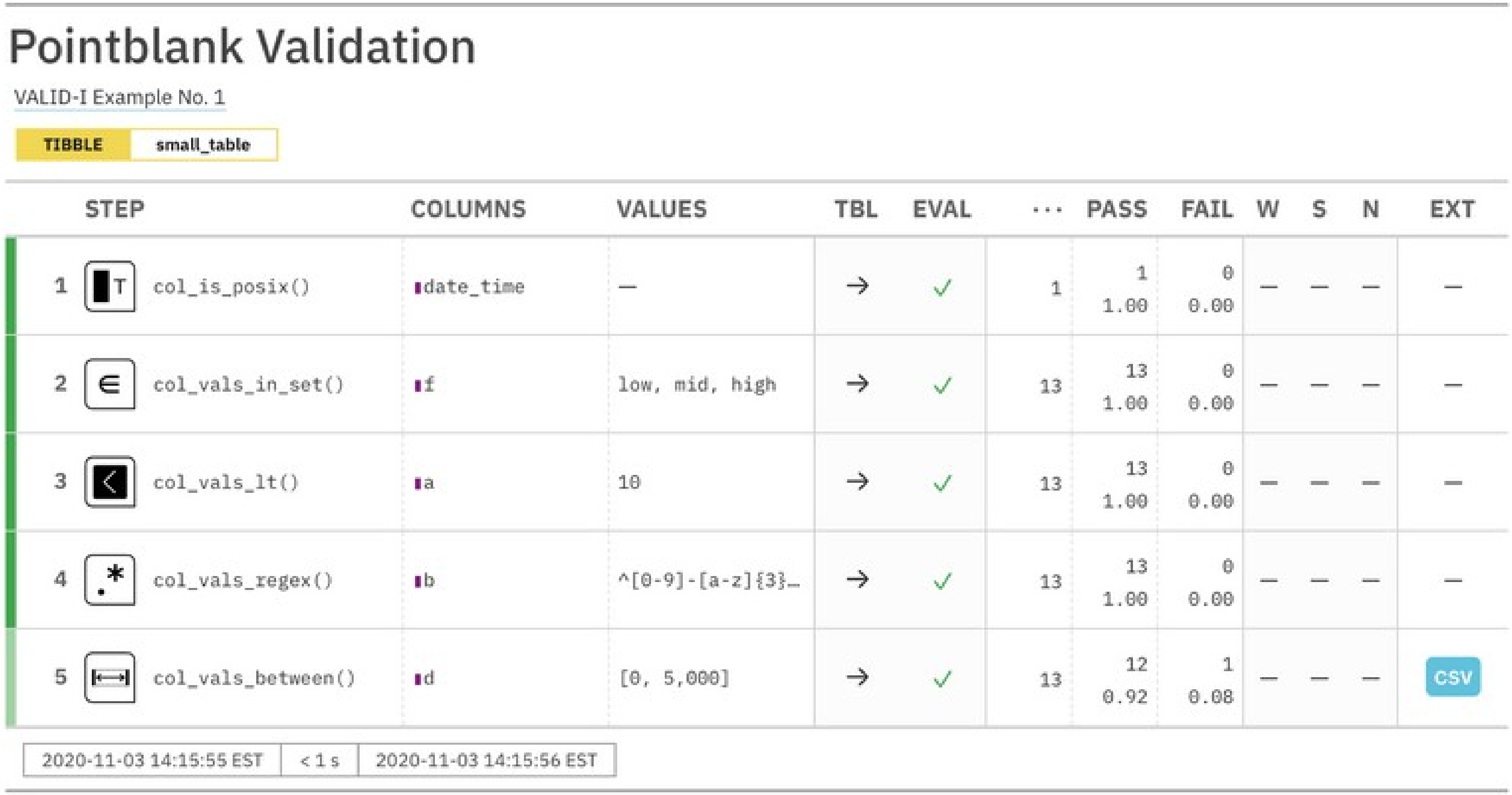


```
── Interrogation Started - there are 5 steps ──────────────────────────────
✓ Step 1: OK.
✓ Step 2: OK.
✓ Step 3: OK.
✓ Step 4: OK.
✓ Step 5: OK.

── Interrogation Completed ────────────────────────────────────────────────
```

From https://rich-iannone.github.io/pointblank/articles/validation_workflows.html

Pointblank : basic example result

```
agent
```



Pointblank Validation

VALID-I Example No. 1

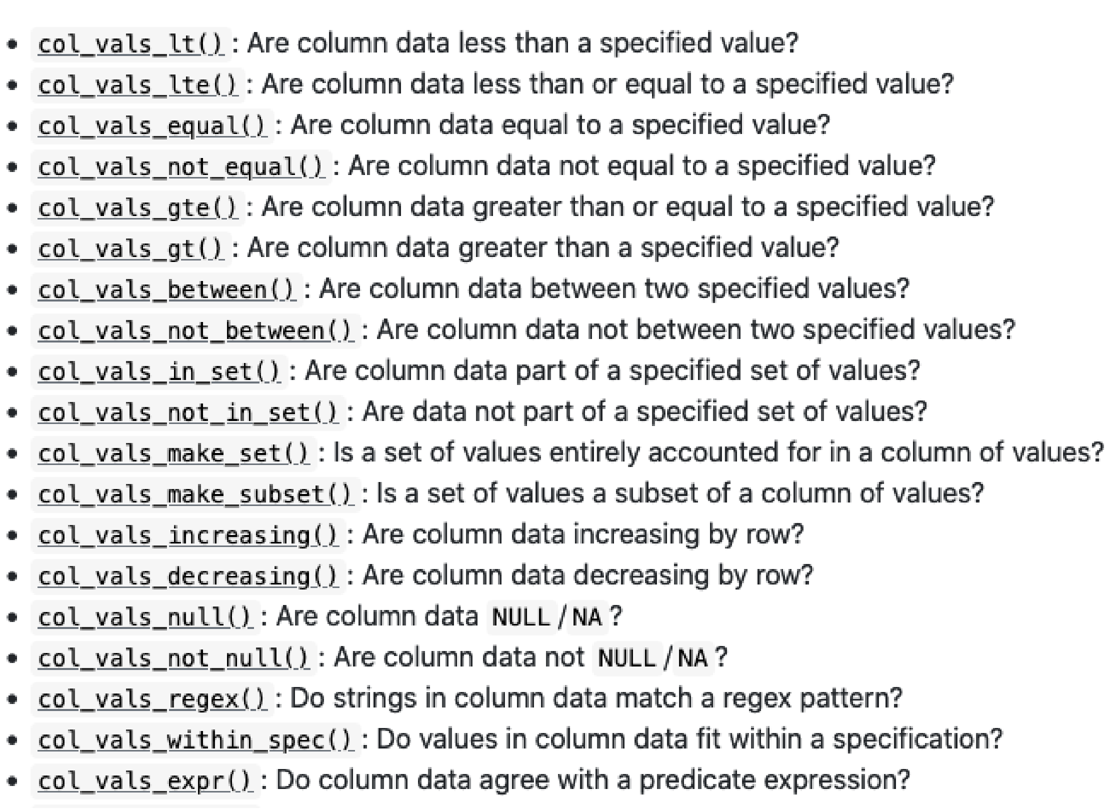
TIBBLE small_table

	STEP	COLUMNS	VALUES	TBL	EVAL	···	PASS	FAIL	W	S	N	EXT
1	▮T col_is_posix()	date_time	—	→	✓	1	1 1.00	0 0.00	—	—	—	—
2	∈ col_vals_in_set()	f	low, mid, high	→	✓	13	13 1.00	0 0.00	—	—	—	—
3	< col_vals_lt()	a	10	→	✓	13	13 1.00	0 0.00	—	—	—	—
4	.* col_vals_regex()	b	^[0-9]-[a-z]{3}…	→	✓	13	13 1.00	0 0.00	—	—	—	—
5	↔ col_vals_between()	d	[0, 5,000]	→	✓	13	12 0.92	1 0.08	—	—	—	CSV

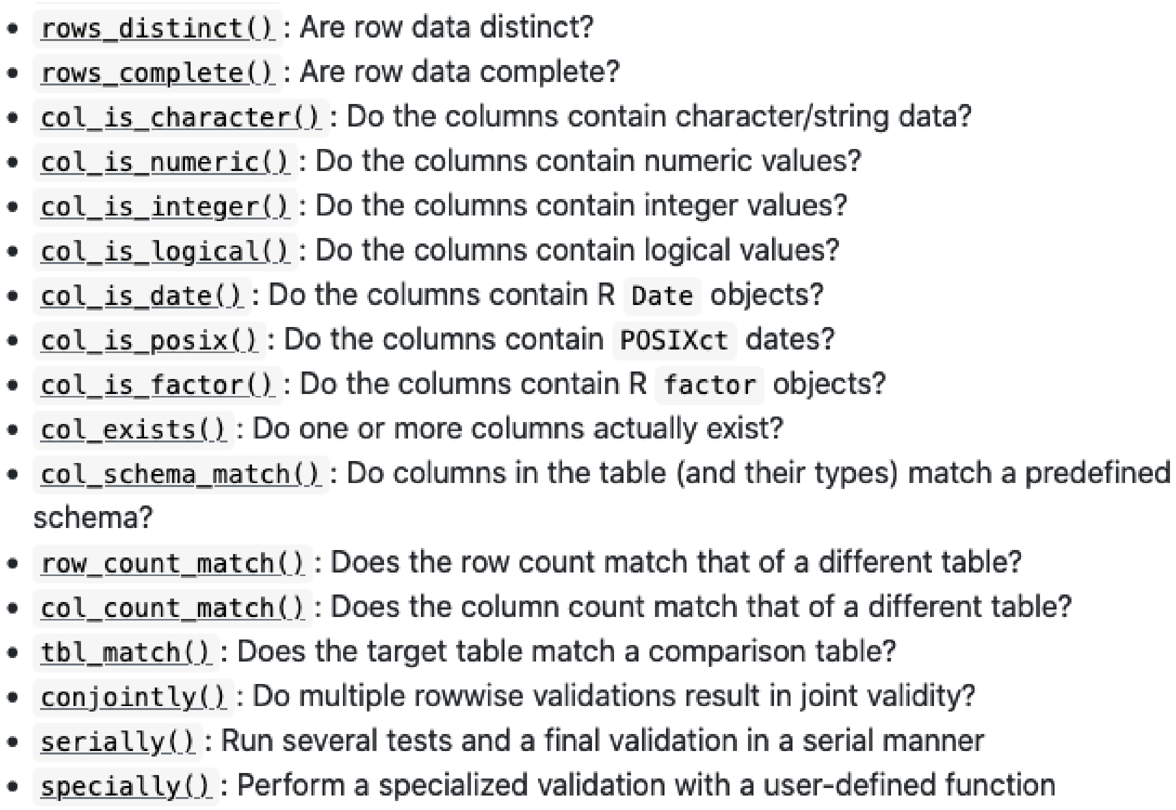
2020-11-03 14:15:55 EST | < 1 s | 2020-11-03 14:15:56 EST

From https://rich-iannone.github.io/pointblank/articles/validation_workflows.html

Pointblank : validation functions



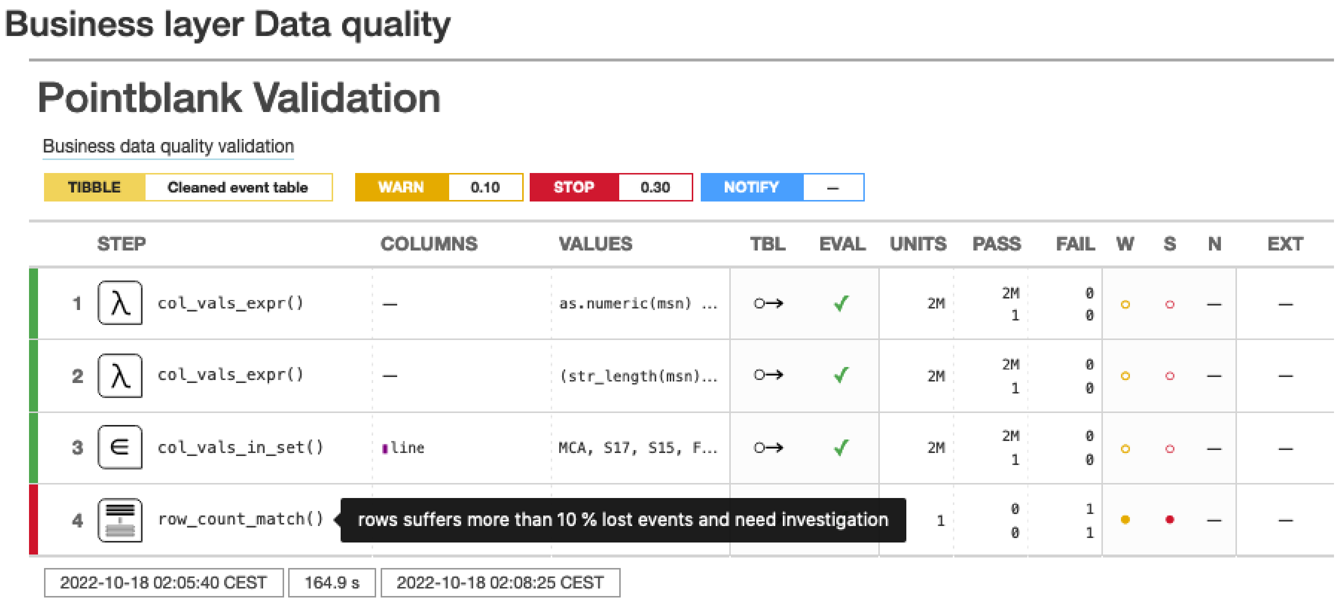
- `col_vals_lt()` : Are column data less than a specified value?
- `col_vals_lte()` : Are column data less than or equal to a specified value?
- `col_vals_equal()` : Are column data equal to a specified value?
- `col_vals_not_equal()` : Are column data not equal to a specified value?
- `col_vals_gte()` : Are column data greater than or equal to a specified value?
- `col_vals_gt()` : Are column data greater than a specified value?
- `col_vals_between()` : Are column data between two specified values?
- `col_vals_not_between()` : Are column data not between two specified values?
- `col_vals_in_set()` : Are column data part of a specified set of values?
- `col_vals_not_in_set()` : Are data not part of a specified set of values?
- `col_vals_make_set()` : Is a set of values entirely accounted for in a column of values?
- `col_vals_make_subset()` : Is a set of values a subset of a column of values?
- `col_vals_increasing()` : Are column data increasing by row?
- `col_vals_decreasing()` : Are column data decreasing by row?
- `col_vals_null()` : Are column data `NULL` / `NA` ?
- `col_vals_not_null()` : Are column data not `NULL` / `NA` ?
- `col_vals_regex()` : Do strings in column data match a regex pattern?
- `col_vals_within_spec()` : Do values in column data fit within a specification?
- `col_vals_expr()` : Do column data agree with a predicate expression?



- `rows_distinct()` : Are row data distinct?
- `rows_complete()` : Are row data complete?
- `col_is_character()` : Do the columns contain character/string data?
- `col_is_numeric()` : Do the columns contain numeric values?
- `col_is_integer()` : Do the columns contain integer values?
- `col_is_logical()` : Do the columns contain logical values?
- `col_is_date()` : Do the columns contain R `Date` objects?
- `col_is_posix()` : Do the columns contain `POSIXct` dates?
- `col_is_factor()` : Do the columns contain R `factor` objects?
- `col_exists()` : Do one or more columns actually exist?
- `col_schema_match()` : Do columns in the table (and their types) match a predefined schema?
- `row_count_match()` : Does the row count match that of a different table?
- `col_count_match()` : Does the column count match that of a different table?
- `tbl_match()` : Does the target table match a comparison table?
- `conjointly()` : Do multiple rowwise validations result in joint validity?
- `serially()` : Run several tests and a final validation in a serial manner
- `specially()` : Perform a specialized validation with a user-defined function

From https://rich-iannone.github.io/pointblank/articles/validation_workflows.html

After poinblank



Business layer Data quality

Pointblank Validation

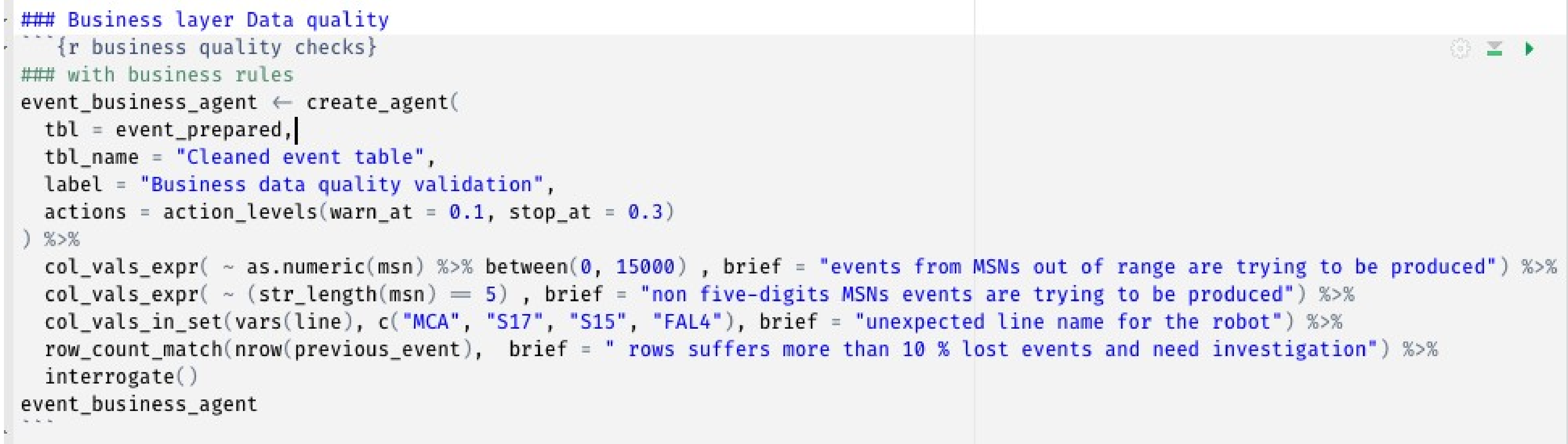
Business data quality validation

TIBBLE | Cleaned event table   WARN | 0.10   STOP | 0.30   NOTIFY | —

	STEP	COLUMNS	VALUES	TBL	EVAL	UNITS	PASS	FAIL	W	S	N	EXT
1	λ col_vals_expr()	—	as.numeric(msn) …	○→	✓	2M	2M 1	0 0	○	○	—	—
2	λ col_vals_expr()	—	(str_length(msn)…	○→	✓	2M	2M 1	0 0	○	○	—	—
3	∈ col_vals_in_set()	▮line	MCA, S17, S15, F…	○→	✓	2M	2M 1	0 0	○	○	—	—
4	row_count_match()					1	0 0	1 1	●	●	—	—

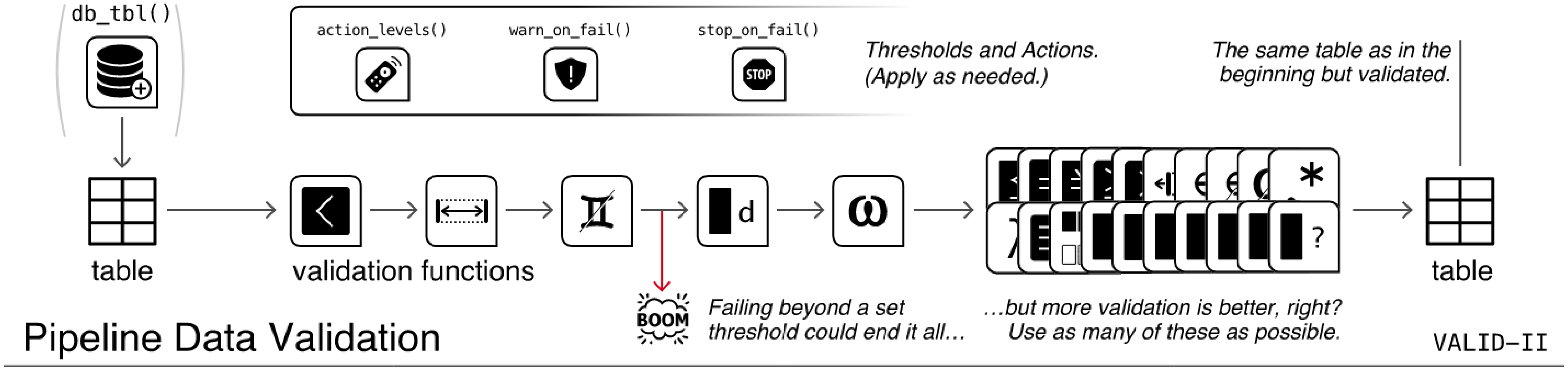
rows suffers more than 10 % lost events and need investigation

2022-10-18 02:05:40 CEST   164.9 s   2022-10-18 02:08:25 CEST



~~~
### Business layer Data quality
```{r business quality checks}
### with business rules
event_business_agent ← create_agent(
  tbl = event_prepared,
  tbl_name = "Cleaned event table",
  label = "Business data quality validation",
  actions = action_levels(warn_at = 0.1, stop_at = 0.3)
) %>%
  col_vals_expr( ~ as.numeric(msn) %>% between(0, 15000) , brief = "events from MSNs out of range are trying to be produced") %>%
  col_vals_expr( ~ (str_length(msn) == 5) , brief = "non five-digits MSNs events are trying to be produced") %>%
  col_vals_in_set(vars(line), c("MCA", "S17", "S15", "FAL4"), brief = "unexpected line name for the robot") %>%
  row_count_match(nrow(previous_event),  brief = " rows suffers more than 10 % lost events and need investigation") %>%
  interrogate()
event_business_agent
```
~~~

Pointblank : Pipeline Validation setup logic



Thresholds and Actions. (Apply as needed.)

The same table as in the beginning but validated.

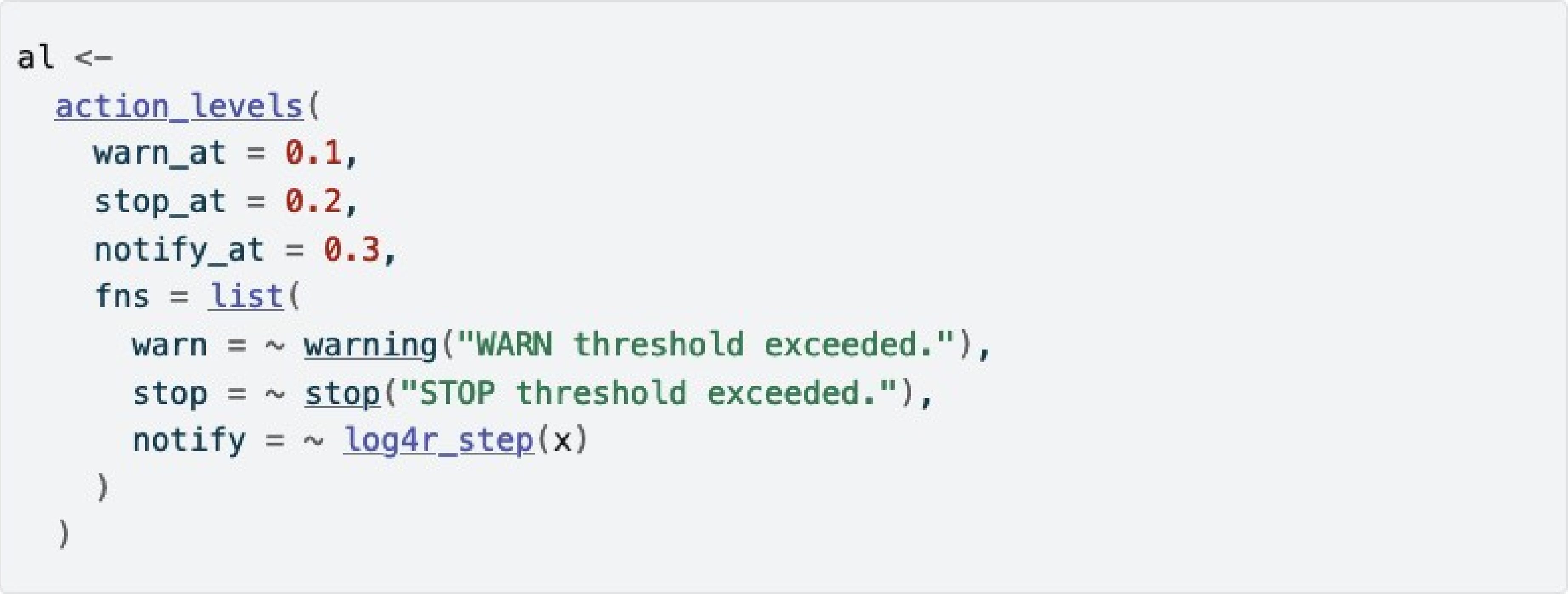
Failing beyond a set threshold could end it all…

…but more validation is better, right? Use as many of these as possible.

Pipeline Data Validation

From https://rich-iannone.github.io/pointblank/articles/validation_workflows.html

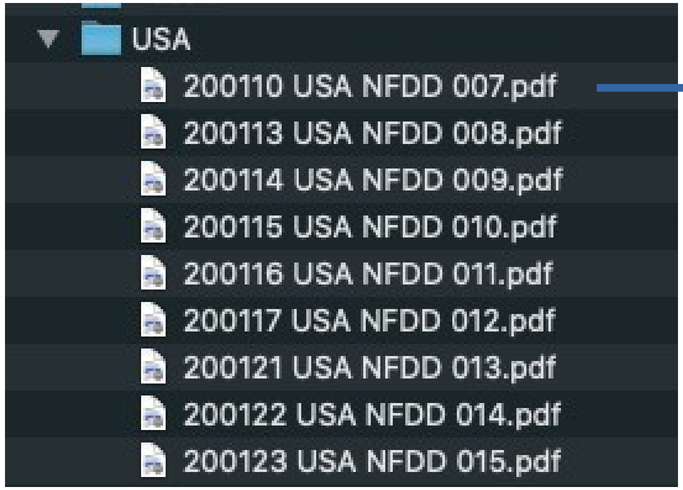
Pointblank : Pipeline Validation setup logic



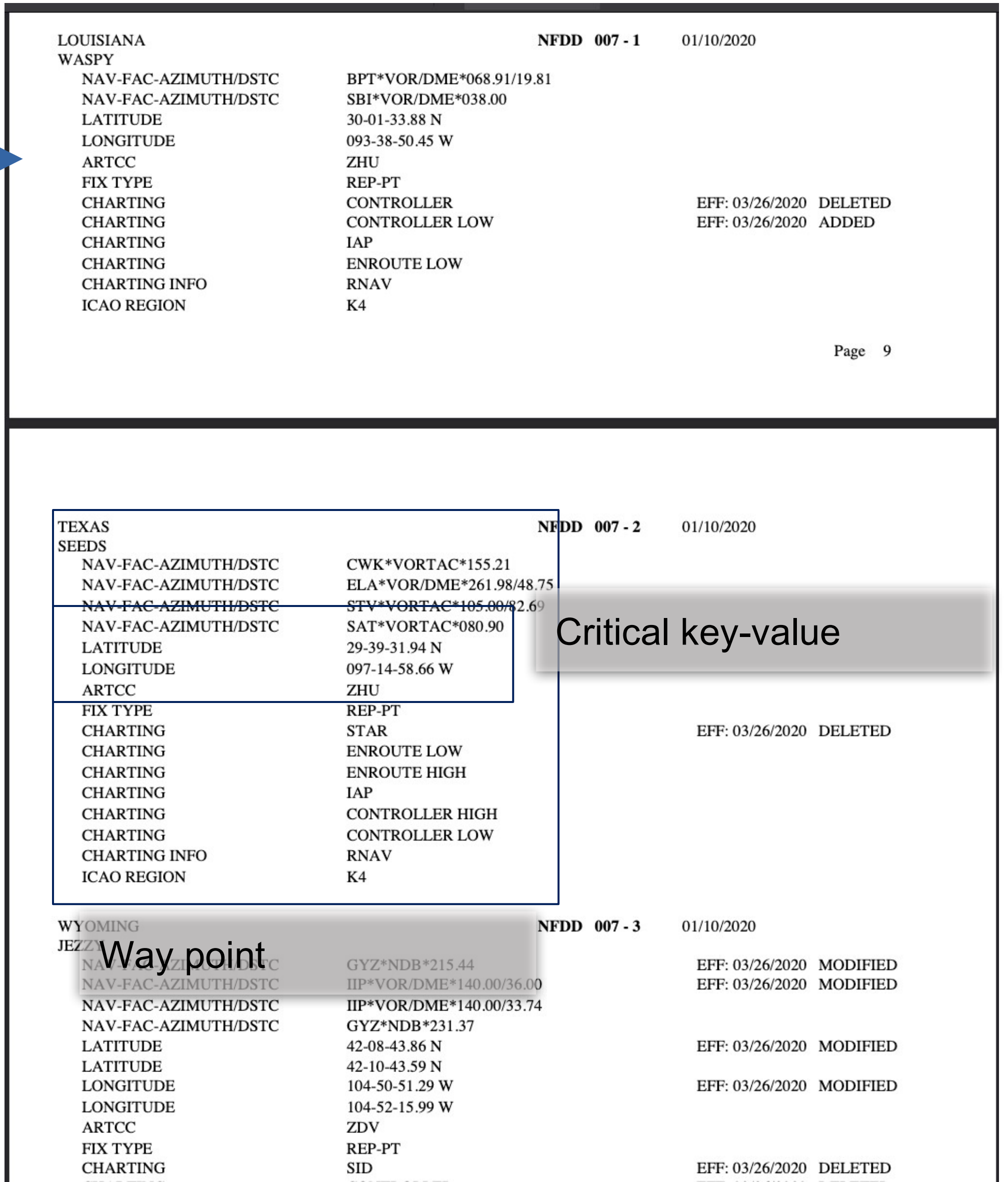
```
al <-
  action_levels(
    warn_at = 0.1,
    stop_at = 0.2,
    notify_at = 0.3,
    fns = list(
      warn = ~ warning("WARN threshold exceeded."),
      stop = ~ stop("STOP threshold exceeded."),
      notify = ~ log4r_step(x)
    )
  )
```

From https://rich-iannone.github.io/pointblank/articles/validation_workflows.html

Pointblank : use it for hackathon !



USA
- 200110 USA NFDD 007.pdf
- 200113 USA NFDD 008.pdf
- 200114 USA NFDD 009.pdf
- 200115 USA NFDD 010.pdf
- 200116 USA NFDD 011.pdf
- 200117 USA NFDD 012.pdf
- 200121 USA NFDD 013.pdf
- 200122 USA NFDD 014.pdf
- 200123 USA NFDD 015.pdf

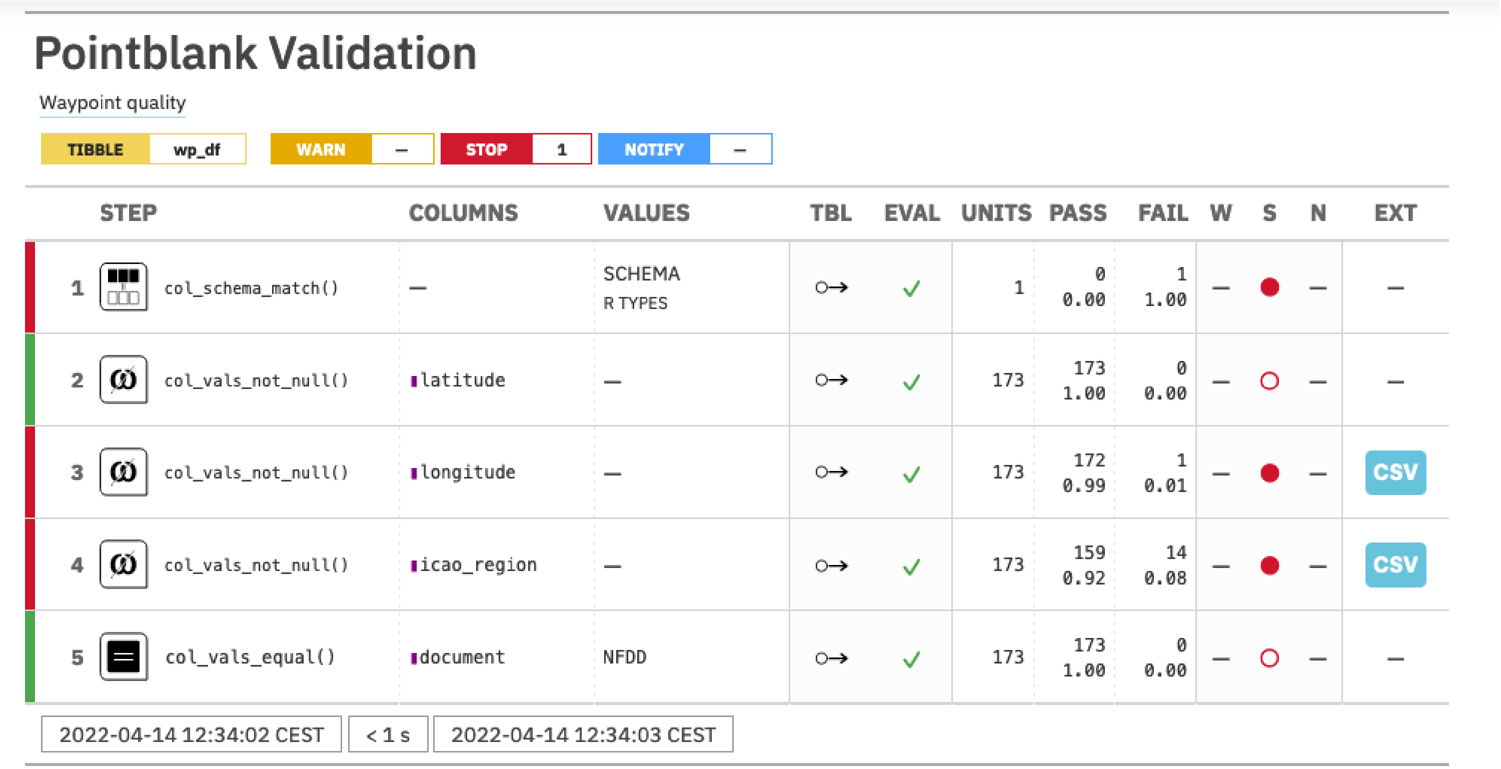


```
LOUISIANA                                         NFDD  007 - 1    01/10/2020
WASPY
  NAV-FAC-AZIMUTH/DSTC        BPT*VOR/DME*068.91/19.81
  NAV-FAC-AZIMUTH/DSTC        SBI*VOR/DME*038.00
  LATITUDE                    30-01-33.88 N
  LONGITUDE                   093-38-50.45 W
  ARTCC                       ZHU
  FIX TYPE                    REP-PT
  CHARTING                    CONTROLLER                            EFF: 03/26/2020  DELETED
  CHARTING                    CONTROLLER LOW                        EFF: 03/26/2020  ADDED
  CHARTING                    IAP
  CHARTING                    ENROUTE LOW
  CHARTING INFO               RNAV
  ICAO REGION                 K4

                                                                               Page   9


TEXAS                                             NFDD  007 - 2    01/10/2020
SEEDS
  NAV-FAC-AZIMUTH/DSTC        CWK*VORTAC*155.21
  NAV-FAC-AZIMUTH/DSTC        ELA*VOR/DME*261.98/48.75
  NAV-FAC-AZIMUTH/DSTC        STV*VORTAC*105.00/82.69
  NAV-FAC-AZIMUTH/DSTC        SAT*VORTAC*080.90
  LATITUDE                    29-39-31.94 N
  LONGITUDE                   097-14-58.66 W
  ARTCC                       ZHU
  FIX TYPE                    REP-PT
  CHARTING                    STAR                                  EFF: 03/26/2020  DELETED
  CHARTING                    ENROUTE LOW
  CHARTING                    ENROUTE HIGH
  CHARTING                    IAP
  CHARTING                    CONTROLLER HIGH
  CHARTING                    CONTROLLER LOW
  CHARTING INFO               RNAV
  ICAO REGION                 K4

WYOMING                                           NFDD  007 - 3    01/10/2020
JEZZY
  NAV-FAC-AZIMUTH/DSTC        GYZ*NDB*215.44                        EFF: 03/26/2020  MODIFIED
  NAV-FAC-AZIMUTH/DSTC        IIP*VOR/DME*140.00/36.00              EFF: 03/26/2020  MODIFIED
  NAV-FAC-AZIMUTH/DSTC        IIP*VOR/DME*140.00/33.74
  NAV-FAC-AZIMUTH/DSTC        GYZ*NDB*231.37
  LATITUDE                    42-08-43.86 N                         EFF: 03/26/2020  MODIFIED
  LATITUDE                    42-10-43.59 N
  LONGITUDE                   104-50-51.29 W                        EFF: 03/26/2020  MODIFIED
  LONGITUDE                   104-52-15.99 W
  ARTCC                       ZDV
  FIX TYPE                    REP-PT
  CHARTING                    SID                                   EFF: 03/26/2020  DELETED
```

Critical key-value

Way point

Pointblank : use it for hackathon !



Pointblank Validation

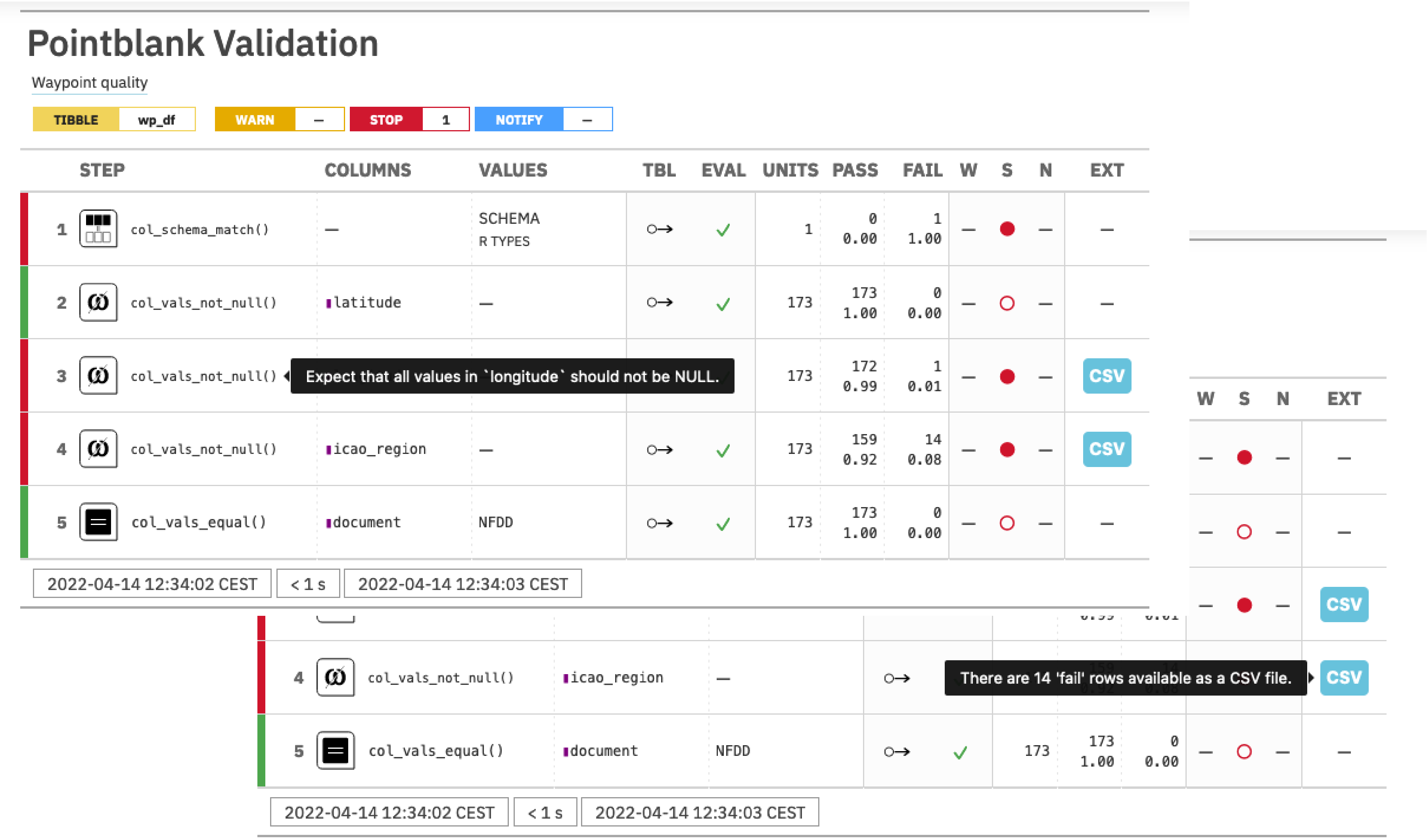
Waypoint quality

TIBBLE wp_df | WARN — | STOP 1 | NOTIFY —

	STEP	COLUMNS	VALUES	TBL	EVAL	UNITS	PASS	FAIL	W	S	N	EXT
1	col_schema_match()	—	SCHEMA R TYPES	○→	✓	1	0 0.00	1 1.00	—	●	—	—
2	col_vals_not_null()	▮latitude	—	○→	✓	173	173 1.00	0 0.00	—	○	—	—
3	col_vals_not_null()	▮longitude	—	○→	✓	173	172 0.99	1 0.01	—	●	—	CSV
4	col_vals_not_null()	▮icao_region	—	○→	✓	173	159 0.92	14 0.08	—	●	—	CSV
5	col_vals_equal()	▮document	NFDD	○→	✓	173	173 1.00	0 0.00	—	○	—	—

2022-04-14 12:34:02 CEST | < 1 s | 2022-04-14 12:34:03 CEST

Pointblank : use it for hackathon !



Pointblank Validation

Waypoint quality

TIBBLE wp_df | WARN — | STOP 1 | NOTIFY —

STEP	COLUMNS	VALUES	TBL	EVAL	UNITS	PASS	FAIL	W	S	N	EXT
1 col_schema_match()	—	SCHEMA R TYPES	○→	✓	1	0 0.00	1 1.00	—	●	—	—
2 col_vals_not_null()	latitude	—	○→	✓	173	173 1.00	0 0.00	—	○	—	—
3 col_vals_not_null()					173	172 0.99	1 0.01	—	●	—	CSV
4 col_vals_not_null()	icao_region	—	○→	✓	173	159 0.92	14 0.08	—	●	—	CSV
5 col_vals_equal()	document	NFDD	○→	✓	173	173 1.00	0 0.00	—	○	—	—

Expect that all values in `longitude` should not be NULL.

2022-04-14 12:34:02 CEST | < 1 s | 2022-04-14 12:34:03 CEST

W	S	N	EXT
—	●	—	—
—	○	—	—
—	●	—	CSV

STEP	COLUMNS	VALUES	TBL	EVAL	UNITS	PASS	FAIL	W	S	N	EXT
4 col_vals_not_null()	icao_region	—	○→					—		—	CSV
5 col_vals_equal()	document	NFDD	○→	✓	173	173 1.00	0 0.00	—	○	—	—

There are 14 'fail' rows available as a CSV file.

2022-04-14 12:34:02 CEST | < 1 s | 2022-04-14 12:34:03 CEST

Pointblank : TDDD ?

Try a new development methodology :
Test-Driven Data Development

Thank you