



Learnings using R for creating CDISC data for Analysis and Reporting purposes!

Gustav Söderström – gustav.soderstrom@astrazeneca.com

Srinivasa Pillarisetty – srinivasa.pillarisetty@astrazeneca.com

Nicole Major – nicole.major@astrazeneca.com

Agenda

1 Background

2 SDTM

3 ADaM

4 Thoughts and reflections



Background

Commitment within AZ Biometrics to build internal R capabilities

Training courses

Replicating work typically done in SAS within R

- Ex: Pilot to create several SDTMs, ADaMs, and TLFs

Starting with internal projects before moving to external projects requiring regulatory approval

Training steps performed

Package development

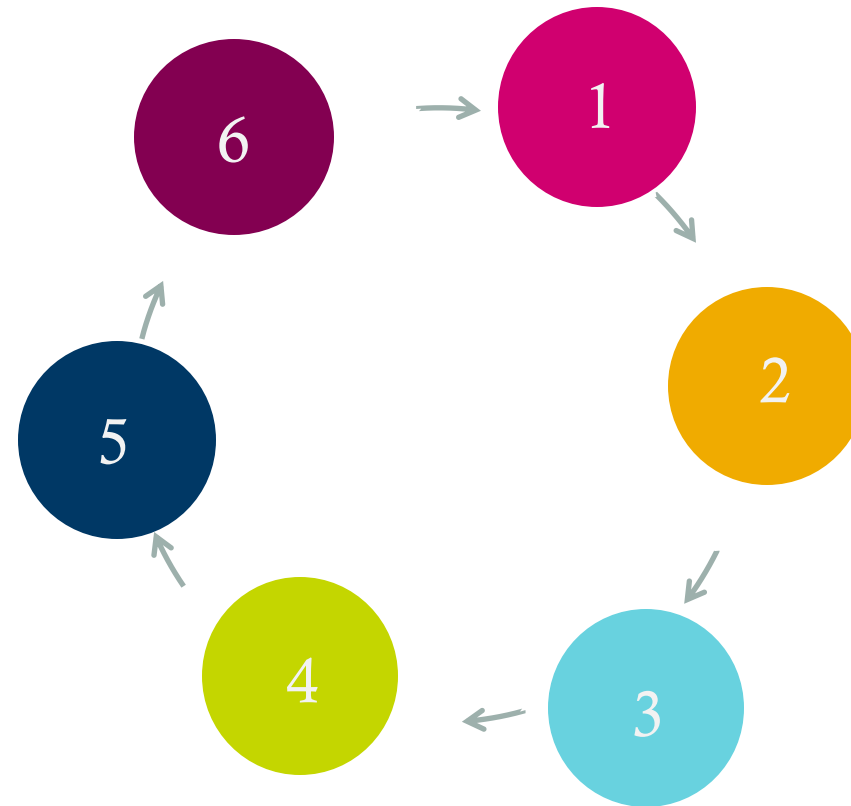
Introduction to functions
and package development

Shiny

Introduction to how to build
interactive applications

R Notebooks

Basics of R Markdown



Tidyverse

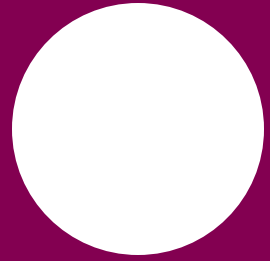
Learned the basics of R using
the Tidyverse package

ggplot

Plotting figures and graphs
using the ggplot2 library

Git & RStudio

Versioning control



SDTM

AE

Set up

Created an R Project to hold all my relevant files (code, datasets, etc.).

Load packages

```
library("tidyverse") # Used for data manipulation and to read in datasets (readr)
library("arsenal")   # Used for comparing datasets
```

Load data

Create formats

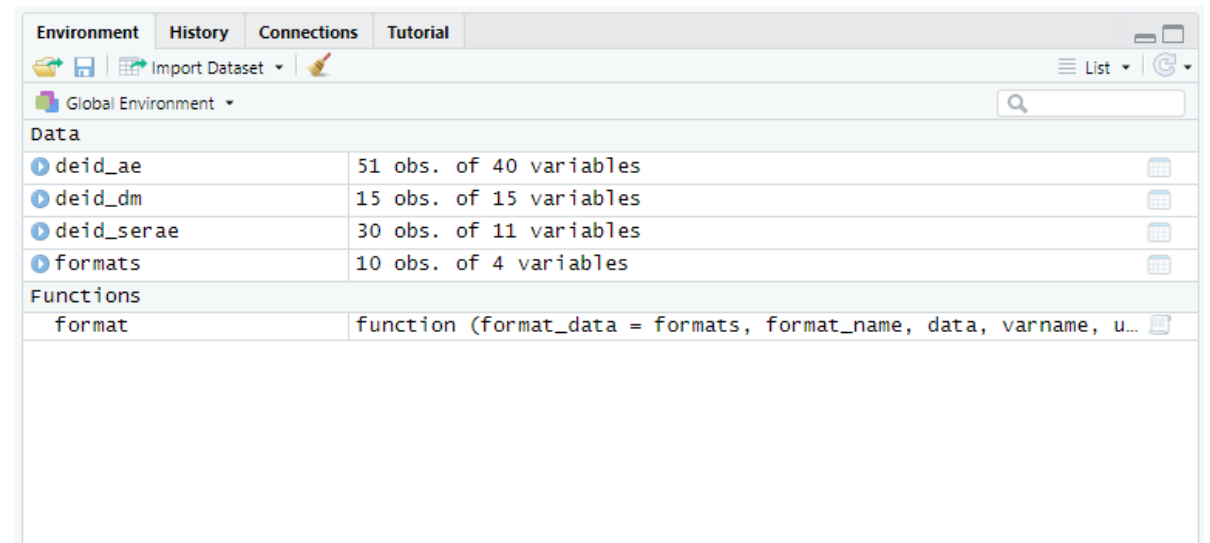
```
formats = data.frame(FMTNAME = c(rep("AEACN", 5)),
                     START   = c("0", "2", "3", "4", "98"),
                     END     = c("0", "2", "3", "4", "98"),
                     LABEL   = c("Dose Not Changed", "Dose Reduced", "Drug Interrupted",
                                "Drug Permanently Discontinued", "Not Applicable"))
```

Source datasets

```
deid_ae <- read_csv(here::here("deid_ae.csv"))
deid_serae <- read_csv(here::here("deid_serae.csv"))
deid_dm <- read_csv(here::here("deid_dm.csv"))
```

Load functions

```
source(here::here("format.R"))
```



Merging

Merge deid_ae and deid_serae to form intermediate AE dataset

```
deid_serae %>%
```

```
  select(-STUDY, -AETERM) -> deid_serae
```

```
deid_ae %>%
```

```
  left_join(deid_serae, by = c("USUBJID", "SUBJECT", "AENO")) -> deid_ae
```

Merge deid_dm to get rfstdtc and rfendtc

```
deid_dm %>%
```

```
  select(USUBJID, RFSTDTC, RFENDTC) %>%
```

```
  right_join(deid_ae, by = "USUBJID") -> ae0
```

```
colnames(ae0)
```

```
[1] "USUBJID" "RFSTDTC" "RFENDTC" "SUBJECT" "STUDY" "AENO" "AETERM" "AESTDAT" "AESER"
```

```
[10] "AEACN" "AEREL" "AETOXGR" "AEC01TOX" "AEC02TOX" "AEC03TOX" "AEC04TOX" "AEC05TOX" "AEC06TOX"
```

```
[19] "AEC07TOX" "AEC08TOX" "AEC09TOX" "AEC01AC" "AEC02AC" "AEC03AC" "AEC04AC" "AEC05AC" "AEC06AC"
```

```
[28] "AEC07AC" "AEC08AC" "AEC09AC" "AEOUT" "AEENDAT" "LLT_CODE" "LLT_NAME" "PT_CODE" "PT_NAME"
```

```
[37] "HLT_CODE" "HLT_NAME" "HLGTCODE" "HLGTNAME" "SOC_CODE" "SOC_NAME" "AESDTH" "AESHOSP" "AESCONG"
```

```
[46] "AESLIFE" "AESDISAB" "AESMIE"
```

deid_serae

deid_ae

deid_dm

AEACN derivation

selecting the needed variables for the derivation

ae0 %>%

```
select(USUBJID, AENO, AETERM, (starts_with("AEC") & ends_with("AC"))) %>%
```

```
pivot_longer(cols = AEC01AC:AEC09AC, names_to = "Column", values_to = "AEAC") %>% # transpose (wide to long format)
```

```
group_by(USUBJID, AENO, AETERM) %>%
```

```
mutate(order = seq(1:n())) %>% # number each action taken within the same USUBJID, AE term, and AE number.
```

```
filter(AEAC %in% c(2, 3, 4)) %>%
```

```
filter(order == max(order)) %>% # get most recent action taken with a value of 2, 3, or 4
```

```
select(USUBJID, AENO, AETERM, AEAC) -> dose_changes
```

USUBJID <chr>	AENO <dbl>	AETERM <chr>	AEC01AC <dbl>	AEC02AC <dbl>
ELX5972/E783131	1	Pneumonia	4	0
ELX5972/E659827	1	Internal Bleeding	NA	NA
ELX5972/E659827	2	Rashes	4	98
ELX5972/E659827	3	Pneumonia	NA	NA
ELX5972/E659827	4	Chills	0	NA
ELX5972/E993872	1	Chills	98	4
ELX5972/E993872	2	Insomnia	3	98
ELX5972/E993872	3	Joint Pain	0	0
ELX5972/E993872	4	Stomachache	0	2
ELX5972/E536888	1	Pneumonia	4	0

Select statement



USUBJID <chr>	AE... <dbl>	AETERM <chr>	Column <chr>	AEAC <dbl>
ELX5972/E783131	1	Pneumonia	AEC01AC	4
ELX5972/E783131	1	Pneumonia	AEC02AC	0
ELX5972/E783131	1	Pneumonia	AEC03AC	NA
ELX5972/E783131	1	Pneumonia	AEC04AC	NA
ELX5972/E783131	1	Pneumonia	AEC05AC	NA
ELX5972/E783131	1	Pneumonia	AEC06AC	NA
ELX5972/E783131	1	Pneumonia	AEC07AC	NA
ELX5972/E783131	1	Pneumonia	AEC08AC	NA
ELX5972/E783131	1	Pneumonia	AEC09AC	NA
ELX5972/E659827	1	Internal Bleeding	AEC01AC	NA

Transpose



USUBJID <chr>	AENO <dbl>	AETERM <chr>	Column <chr>	AEAC <dbl>	order <int>
ELX5972/E783131	1	Pneumonia	AEC01AC	4	1
ELX5972/E783131	1	Pneumonia	AEC02AC	0	2
ELX5972/E783131	1	Pneumonia	AEC03AC	NA	3
ELX5972/E783131	1	Pneumonia	AEC04AC	NA	4
ELX5972/E783131	1	Pneumonia	AEC05AC	NA	5
ELX5972/E783131	1	Pneumonia	AEC06AC	NA	6
ELX5972/E783131	1	Pneumonia	AEC07AC	NA	7
ELX5972/E783131	1	Pneumonia	AEC08AC	NA	8
ELX5972/E783131	1	Pneumonia	AEC09AC	NA	9
ELX5972/E659827	1	Internal Bleeding	AEC01AC	NA	1

Mutate + filter statement



USUBJID <chr>	AENO <dbl>	AETERM <chr>	Column <chr>	AEAC <dbl>	order <int>
ELX5972/E783131	1	Pneumonia	AEC01AC	4	1
ELX5972/E659827	2	Rashes	AEC01AC	4	1
ELX5972/E993872	1	Chills	AEC07AC	2	7
ELX5972/E993872	2	Insomnia	AEC07AC	4	7
ELX5972/E993872	3	Joint Pain	AEC08AC	3	8
ELX5972/E993872	4	Stomachache	AEC08AC	3	8
ELX5972/E536888	1	Pneumonia	AEC03AC	4	3
ELX5972/E536888	2	Diarrhea	AEC02AC	2	2
ELX5972/E566930	2	Internal Bleeding	AEC01AC	4	1
ELX5972/E566930	3	Nausea	AEC07AC	3	7

Filter by numbers + max(order)

AEACN derivation (cont.)

update value for AEACN if AEAC is not missing

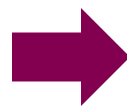
ae0 %>%

left_join(dose_changes, by = c("USUBJID", "AENO", "AETERM")) %>%

mutate(AEACN = if_else(!is.na(AEAC), AEAC, AEACN)) -> ae1

USUBJID <chr>	AENO <dbl>	AETERM <chr>	AEAC <dbl>
ELX5972/E783131	1	Pneumonia	4
ELX5972/E659827	2	Rashes	4
ELX5972/E993872	1	Chills	2
ELX5972/E993872	2	Insomnia	4
ELX5972/E993872	3	Joint Pain	3
ELX5972/E993872	4	Stomachache	3
ELX5972/E536888	1	Pneumonia	4
ELX5972/E536888	2	Diarrhea	2
ELX5972/E566930	2	Internal Bleeding	4
ELX5972/E566930	3	Nausea	3

dose_changes



USUBJID <chr>	AENO <dbl>	AETERM <chr>	AEACN <dbl>
ELX5972/E783131	1	Pneumonia	4
ELX5972/E659827	1	Internal Bleeding	0
ELX5972/E659827	2	Rashes	4
ELX5972/E659827	3	Pneumonia	2
ELX5972/E659827	4	Chills	2
ELX5972/E993872	1	Chills	2
ELX5972/E993872	2	Insomnia	4
ELX5972/E993872	3	Joint Pain	3
ELX5972/E993872	4	Stomachache	3
ELX5972/E536888	1	Pneumonia	4

ae1

Formatting variables

ae1 %>%

```
mutate(AEACN = format(format_name = "AEACN", data = ., varname = "AEACN")) -> ae2
```

USUBJID <chr>	AENO <dbl>	AETERM <chr>	AEACN <dbl>
ELX5972/E783131	1	Pneumonia	4
ELX5972/E659827	1	Internal Bleeding	0
ELX5972/E659827	2	Rashes	4
ELX5972/E659827	3	Pneumonia	2
ELX5972/E659827	4	Chills	2
ELX5972/E993872	1	Chills	2
ELX5972/E993872	2	Insomnia	4
ELX5972/E993872	3	Joint Pain	3
ELX5972/E993872	4	Stomachache	3
ELX5972/E536888	1	Pneumonia	4

ae1



USUBJID <chr>	AENO <dbl>	AETERM <chr>	AEACN <chr>
ELX5972/E783131	1	Pneumonia	DRUG PERMANENTLY DISCONTINUED
ELX5972/E659827	1	Internal Bleeding	DOSE NOT CHANGED
ELX5972/E659827	2	Rashes	DRUG PERMANENTLY DISCONTINUED
ELX5972/E659827	3	Pneumonia	DOSE REDUCED
ELX5972/E659827	4	Chills	DOSE REDUCED
ELX5972/E993872	1	Chills	DOSE REDUCED
ELX5972/E993872	2	Insomnia	DRUG PERMANENTLY DISCONTINUED
ELX5972/E993872	3	Joint Pain	DRUG INTERRUPTED
ELX5972/E993872	4	Stomachache	DRUG INTERRUPTED
ELX5972/E536888	1	Pneumonia	DRUG PERMANENTLY DISCONTINUED

Formatted AEACN

You could also use factors to accomplish the same thing.

AESEQ derivation

```
ae2 %>%  
  group_by(USUBJID) %>%  
  arrange(AEDECOD, AESTDTC, AEENDTC, AESPID, .by_group = TRUE) %>%  
  mutate(AESEQ = as.numeric(seq(1:n())))) -> ae3
```

USUBJID <chr>	AESEQ <dbl>	AETERM <chr>	AEDECOD <chr>	AESTDTC <chr>	AEENDTC <chr>	AESPID <chr>
ELX5972/E121809	1	Joint Pain	Arthralgia	2020-11-28	2021-01-03	2
ELX5972/E121809	2	Chills	Chills	2020-11-26	2021-01-20	5
ELX5972/E121809	3	Cough	Cough	2020-10-20	2020-10-21	4
ELX5972/E121809	4	Nausea	Nausea	2021-01-07	2021-02-09	1
ELX5972/E121809	5	Pneumonia	Pneumonia	2021-03-04	2021-04-25	3
ELX5972/E157053	1	Internal Bleeding	Internal haemorrhage	2021-01-28	2021-02-04	1
ELX5972/E157053	2	Fever	Pyrexia	2021-01-14	2021-02-22	3
ELX5972/E157053	3	Rashes	Rash	2020-11-08	2020-12-05	2
ELX5972/E231017	1	Joint Pain	Arthralgia	2021-01-17	2021-02-08	4
ELX5972/E231017	2	Chills	Chills	2020-12-03	2021-01-21	2

Arsenal package

```
sink("difference_in_sorts")
summary(comparedf(ae4_diff, ae3))
sink()
```

If there's an interest in seeing a summary of the differences as well as the differences themselves, then use this code:

```
summary(comparedf(ae4_diff, ae3))
```

If there's only an interest in seeing the differences, then use this code:

```
diffs(comparedf(ae4_diff, ae3))
```

The sink function is a method of exporting output into a file.

Table: Summary of overall comparison

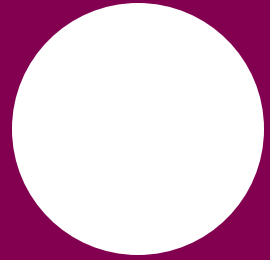
statistic	value
Number of by-variables	0
Number of non-by variables in common	34
Number of variables compared	34
Number of variables in x but not y	0
Number of variables in y but not x	0
Number of variables compared with some values unequal	32
Number of variables compared with all values equal	2
Number of observations in common	51
Number of observations in x but not y	0
Number of observations in y but not x	0
Number of observations with some compared variables unequal	50
Number of observations with all compared variables equal	1
Number of values unequal	1228

Table: Differences detected by variable

var.x	var.y	n	NAs
STUDYID	STUDYID	0	0
DOMAIN	DOMAIN	0	0
USUBJID	USUBJID	49	0
AESEQ	AESEQ	43	0
AESPID	AESPID	38	0
AETERM	AETERM	44	0
AELLT	AELLT	44	0
AELLTCD	AELLTCD	44	0
AEDECOD	AEDECOD	44	0
AEPTCD	AEPTCD	44	0
AEHLT	AEHLT	44	0
AEHLTCD	AEHLTCD	44	0
AEHLGT	AEHLGT	43	0
AEHLGTC	AEHLGTC	43	0
AEBO	AEBO	43	0
AEBO	AEBO	43	0

Table: Differences detected (1178 not shown)

var.x	var.y	..row.names..	values.x	values.y	row.x	row.y
USUBJID	USUBJID	1	ELX5972/E231017	ELX5972/E121809	1	1
USUBJID	USUBJID	2	ELX5972/E370884	ELX5972/E121809	2	2
USUBJID	USUBJID	3	ELX5972/E566930	ELX5972/E121809	3	3
USUBJID	USUBJID	5	ELX5972/E354060	ELX5972/E121809	5	5
USUBJID	USUBJID	6	ELX5972/E354060	ELX5972/E157053	6	6
USUBJID	USUBJID	7	ELX5972/E370884	ELX5972/E157053	7	7
USUBJID	USUBJID	8	ELX5972/E536888	ELX5972/E157053	8	8
USUBJID	USUBJID	9	ELX5972/E536888	ELX5972/E231017	9	9
USUBJID	USUBJID	10	ELX5972/E566930	ELX5972/E231017	10	10
USUBJID	USUBJID	11	ELX5972/E566930	ELX5972/E231017	11	11
AESEQ	AESEQ	1	4	1	1	1
AESEQ	AESEQ	2	5	2	2	2
AESEQ	AESEQ	3	5	3	3	3
AESEQ	AESEQ	5	2	5	5	5
AESEQ	AESEQ	6	3	1	6	6
AESEQ	AESEQ	7	3	2	7	7
AESEQ	AESEQ	8	1	3	8	8
AESEQ	AESEQ	9	2	1	9	9
AESEQ	AESEQ	10	1	2	10	10
AESEO	AESEO	11	4	3	11	11



ADaM

ADSL

Packages used

```
library(haven)
```

```
library(tidyverse)
```

```
library(diffrdf)
```

```
dm <- read_sas( '.../dm.sas7bdat', NULL)
```

```
ie <- read_sas( '.../ie.sas7bdat', NULL)
```

Some Useful Functions

#str function

str(dm)

```
## spec_tbl_df[,16] [15 x 16] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ X1      : num [1:15] 1 2 3 4 5 6 7 8 9 10 ...
## $ USUBJID: chr [1:15] "ELX5972/E783131" "ELX5972/E659827"
"ELX5972/E993872" "ELX5972/E536888" ...
## $ SUBJID : chr [1:15] "E783131" "E659827" "E993872" "E536888" ...
## $ RFSTDTC: Date[1:15], format: "2020-06-30" NA ...
## $ RFENDTC: Date[1:15], format: "2020-10-15" NA ...
## $ RFICDTC: Date[1:15], format: "2020-06-15" "2020-11-26" ...
## $ DTHDTC : Date[1:15], format: "2020-11-13" NA ...
## $ DTHFL  : chr [1:15] "Y" NA NA NA ...
## $ AGE    : num [1:15] 35 55 36 25 33 74 42 18 59 36 ...
## $ SEX    : chr [1:15] "M" "F" "F" "M" ...
## $ RACE   : chr [1:15] "BLACK OR AFRICAN AMERICAN" "OTHER" "AMERICAN
INDIAN OR ALASKA NATIVE" "OTHER" ...
## $ ETHNIC : chr [1:15] "HISPANIC OR LATINO" "NOT HISPANIC OR LATINO" "NOT
-HISPANIC OR LATINO" "NOT HISPANIC OR LATINO" ...
## $ ARM    : chr [1:15] "Placebo" "Screen Failure" "Elixir" "Elixir" ...
## $ ARMCD  : chr [1:15] "PLCB" "SCRNFAIL" "ELXR" "ELXR" ...
## $ COUNTRY: chr [1:15] "MEX" "TUR" "USA" "RUS" ...
## $ REGION : chr [1:15] "Latin America" "Europe" "North America" "Europe"
```

Some Useful Functions (cont.)

#table function

```
table(dm$RACE)
```

```
##  
## AMERICAN INDIAN OR ALASKA NATIVE                ASIAN  
##                                2                2  
##      BLACK OR AFRICAN AMERICAN                CAUCASIAN  
##                                4                2  
##                                OTHER  
##                                5
```

#dim function

```
dim(dm)
```

```
## [1] 15 16
```

Some Useful Functions (cont.)

#slice function

```
ie %>%  
  select(USUBJID) %>%  
  group_by(USUBJID) %>%  
  slice(1) %>%  
  mutate(iefl='Y') -> ie1
```

#paste function

```
dm2 %>%  
  mutate(BRTHDTC1 = ifelse(str_length(BRTHDTC)==4, paste(BRTHDTC, '-01-01', sep=""),  
                           ifelse(str_length(BRTHDTC)==7, paste(BRTHDTC, '-01', sep=' '), BRTHDTC))) %>%  
  mutate(birthdt = as.Date(BRTHDTC1)) %>%  
  mutate(AGE_R = as.integer((RANDDT - birthdt + 1)/365.25)) %>%  
  mutate(AGEGRPN1 = ifelse(AGE_R < 65, 1, 0)) -> dm3
```

Custom function

#last contact date function

```
last_date <- function(tbl, varname=REF){  
  tbl %>%  
    select(USUBJID, {{varname}}) %>%  
    filter(!is.na({{varname}})) %>%  
    group_by(USUBJID) %>%  
    summarise(max_date = max({{varname}}))  
}
```

eg %>%

```
last_date(EGDTC) -> eg_last
```

lb %>%

```
last_date(LBDTC) -> lb_last
```

Diffdf package (Similar to Proc compare)

#first select variables to be compared

```
names_adsl <- names(adsl)
```

```
adsl_val <- adsl_val[, names_adsl]
```

#sort

```
adsl_val <- adsl_val %>% arrange(USUBJID)
```

compare datasets, output compare result to output folder

```
issues <- diffdf(adsl, adsl_val, file = ".../output/adsl")
```

Comparison Results (cont.)

#Differences

Differences found between the objects!

A summary is given below.

Not all Values Compared Equal
All rows are shown in table below

=====	
Variable	No of Differences

RANXL	3
HISTN	8

Comparison Results (cont.)

#Shows the row numbers with mis-matches and values in base and compare dataset

All rows are shown in table below

=====			
VARIABLE	..ROWNUMBER..	BASE	COMPARE

RANXL	348	Other	Eligibility criteria not fulfi...
RANXL	400	Other	Eligibility criteria not fulfi...
RANXL	698	Other	Eligibility criteria not fulfi...

#When all issues are resolved, we see below message

No issues were found!

Thoughts and reflections

- 1 Advantage over SAS in ease of data visualization and plot customization
- 2 Interactive reporting capabilities (R Shiny)
- 3 Can be used as a complement to SAS
- 4 Keep in mind that R is case sensitive
- 5 Need to consider package versions and validation

Thank you!

Feel free to contact us with follow-up questions:

gustav.soderstrom@astrazeneca.com
srinivasa.pillarisetty@astrazeneca.com
nicole.major@astrazeneca.com