Caso2 - Envíos

Sebastian Matias Romero Davila

2024-07-10

Contents

Base de datos inicial

Cargamos librerías a emplear

```
library(tidyverse)
library(rio)
library(kableExtra)
```

Base: base_envíos

```
base_envíos <- import("base_envíos.csv")
head(base_envíos,10) %>%
kable(booktabs = TRUE,format = "latex") %>%
  kable_styling(latex_options = c("striped","scale_down","HOLD_position"))
```

id_envío	origen	destino	fecha_envío	fecha_entrega	monto_envío
45	Arequipa	Trujillo	2022-10-11	19689	62.21
12	Lima	Trujillo	2022-01-21	19182	847.67
42	Piura	Trujillo	2022-12-19	19516	797.43
26	Piura	Trujillo	2023-02-07	19051	360.84
99	Lima	Piura	2022-03-10	19210	680.95
37	Cusco	Arequipa	2022-08-04	19393	145.89
100	Trujillo	Trujillo	2023-12-17	19038	456.54
43	Cusco	Trujillo	2022-07-02	19061	194.23
67	Cusco	Cusco	2022-03-29	19523	431.59
-70	Arequipa	Lima	2022-04-23	19041	645.28

Características Inherentes

Exactitud

Para la representación correcta de la realidad de los datos, empezamos identificando los tipos de variables

```
base_envíos %>% glimpse()
```

```
## Rows: 150
## Columns: 6
## $ id_envío      <int> 45, 12, 42, 26, 99, 37, 100, 43, 67, -70, 52, 47, 48, 6,~
## $ origen        <chr> "Arequipa", "Lima", "Piura", "Piura", "Lima", "Cusco", "~
## $ destino       <chr> "Trujillo", "Trujillo", "Trujillo", "Trujillo", "Piura",~
## $ fecha_envío   <chr> "2022-10-11", "2022-01-21", "2022-12-19", "2023-02-07", ~
## $ fecha_entrega <int> 19689, 19182, 19516, 19051, 19210, 19393, 19038, 19061, ~
## $ monto_envío   <dbl> 62.21, 847.67, 797.43, 360.84, 680.95, 145.89, 456.54, 1~
```

Luego, verificamos si existen inexactitudes

- id_envío: Se encuentran valores enteros negativos
- fecha_envío y fecha_entrega: Formato inadecuado

```r
envinexactos <- base_envíos %>% filter(
  id_envío<0|substr(fecha_envío,3,3) == "-"
)
head(envinexactos,10) %>%
kable(booktabs = TRUE,format = "latex") %>%
  kable_styling(latex_options = c("striped", "scale_down", "HOLD_position"))
```

id_envío	origen	destino	fecha_envío	fecha_entrega	monto_envío
-70	Arequipa	Lima	2022-04-23	19041	645.28
-39	Arequipa	Piura	31-12-2023	19593	251.41
-59	Trujillo	Lima	2022-08-13	19497	NA
-74	NA	Lima	31-12-2023	19615	720.15
-22	Arequipa	Cusco	2022-05-22	NA	775.83
-70	Arequipa	Lima	31-12-2023	19294	NA
-1	NA	Trujillo	2022-04-16	19194	392.12
-34	Trujillo	NA	31-12-2023	19211	336.53
-70	NA	Cusco	2022-12-21	19354	NA
-23	Arequipa	NA	31-12-2023	NA	519.09

Completitud

En la tabla se observan valores faltantes en varias de las filas.

```r
envios_faltantes <- base_envíos %>%
  filter(
    if_any(everything(), is.na)
  )
head(envios_faltantes,10) %>%
kable(booktabs = TRUE,format = "latex") %>%
  kable_styling(latex_options = c("striped", "scale_down", "HOLD_position"))
```

id_envío	origen	destino	fecha_envío	fecha_entrega	monto_envío
52	Arequipa	NA	2023-07-29	19701	50.82
48	Cusco	NA	2022-03-26	19661	311.18
29	NA	Piura	2023-05-05	19044	NA
24	NA	Piura	2023-09-12	19508	203.06
61	Arequipa	NA	2023-03-05	19574	909.37
59	NA	NA	2022-10-09	NA	548.43
90	NA	Arequipa	2023-10-10	19153	556.67
72	Arequipa	NA	2022-03-16	19157	129.95
32	NA	Piura	2023-07-04	19180	344.51
-59	Trujillo	Lima	2022-08-13	19497	NA

Cantidad total de datos faltantes por cada columna en el conjunto de datos.

```r
na_envíos <- colSums(is.na(base_envíos))
na_envíos
```

```
##      id_envío         origen        destino   fecha_envío fecha_entrega
##             0             32             23             0             6
##   monto_envío
##            10
```

Entonces, ajustaremos los valores faltantes (NA) de la siguiente manera:

origen: Se reemplazará con "No identificado".

destino: Se reemplazará con "No identificado".

monto_envío: Se asignará como -1 en caso de estar ausente.

```
base1 <- base_envíos %>% mutate(
  origen = case_when(
    !is.na(origen) ~ origen,
    is.na(origen) ~ "No identificado"
  ),
  destino = case_when(
    !is.na(destino) ~ destino,
    is.na(destino) ~ "No identificado"
  ),
  fecha_entrega = case_when(
    !is.na(fecha_entrega) ~ fecha_entrega,
    is.na(fecha_entrega) ~ NA
  ),
  monto_envío = case_when(
    !is.na(monto_envío) ~ monto_envío,
    is.na(monto_envío) ~ -1
  )
)
head(base1,10) %>%
kable(booktabs = TRUE,format = "latex") %>%
  kable_styling(latex_options = c("striped", "scale_down", "HOLD_position"))
```

id_envío	origen	destino	fecha_envío	fecha_entrega	monto_envío
45	Arequipa	Trujillo	2022-10-11	19689	62.21
12	Lima	Trujillo	2022-01-21	19182	847.67
42	Piura	Trujillo	2022-12-19	19516	797.43
26	Piura	Trujillo	2023-02-07	19051	360.84
99	Lima	Piura	2022-03-10	19210	680.95
37	Cusco	Arequipa	2022-08-04	19393	145.89
100	Trujillo	Trujillo	2023-12-17	19038	456.54
43	Cusco	Trujillo	2022-07-02	19061	194.23
67	Cusco	Cusco	2022-03-29	19523	431.59
-70	Arequipa	Lima	2022-04-23	19041	645.28

Consistencia

Previamente, se identificaron inconsistencias en las variables id_vendedor y nombre_vendedor, así como un problema con el formato de la variable fecha_venta. Para abordar estas cuestiones, se realizaron las siguientes acciones:

```
base2 <- base1 %>% mutate(
  id_envío = case_when(
    id_envío<0 ~ as.integer((-1)*id_envío),
    id_envío>=0 ~ as.integer(id_envío)
    ),
  origen = as.character(origen),
  fecha_envío = as.Date(case_when(
    substr(fecha_envío,3,3) == "-" ~ paste( substr(fecha_envío,7,10),
                                            substr(fecha_envío,4,5),
                                            substr(fecha_envío,1,2),
                                            sep = "-"),
    substr(fecha_envío,3,3) != "-" ~ fecha_envío)
    ),
  fecha_entrega = as.Date(fecha_entrega, origin = "1970-01-01"),
  monto_envío = round(monto_envío,2),
  )
head(base2,10) %>%
kable(booktabs = TRUE,format = "latex") %>%
  kable_styling(latex_options = c("striped", "scale_down", "HOLD_position"))
```

id_envío	origen	destino	fecha_envío	fecha_entrega	monto_envío
45	Arequipa	Trujillo	2022-10-11	2023-11-28	62.21
12	Lima	Trujillo	2022-01-21	2022-07-09	847.67
42	Piura	Trujillo	2022-12-19	2023-06-08	797.43
26	Piura	Trujillo	2023-02-07	2022-02-28	360.84
99	Lima	Piura	2022-03-10	2022-08-06	680.95
37	Cusco	Arequipa	2022-08-04	2023-02-05	145.89
100	Trujillo	Trujillo	2023-12-17	2022-02-15	456.54
43	Cusco	Trujillo	2022-07-02	2022-03-10	194.23
67	Cusco	Cusco	2022-03-29	2023-06-15	431.59
70	Arequipa	Lima	2022-04-23	2022-02-18	645.28

Hay fechas de entrega que están situadas antes de la fecha de envío, procedemos a hacer el intercambio respectivo.

```
base3 <- base2 %>% mutate(
  fecha_temporal = fecha_envío,
    fecha_envío = case_when(
      fecha_envío > fecha_entrega ~ fecha_entrega,
      TRUE ~ fecha_envío
    ),
    fecha_entrega = case_when(
      fecha_temporal > fecha_entrega ~ fecha_temporal,
      is.na(fecha_entrega) ~ NA,
      TRUE ~ fecha_entrega
    ),
) %>% select(-fecha_temporal)
head(base3,10) %>%
kable(booktabs = TRUE,format = "latex") %>%
  kable_styling(latex_options = c("striped", "scale_down", "HOLD_position"))
```

id_envío	origen	destino	fecha_envío	fecha_entrega	monto_envío
45	Arequipa	Trujillo	2022-10-11	2023-11-28	62.21
12	Lima	Trujillo	2022-01-21	2022-07-09	847.67
42	Piura	Trujillo	2022-12-19	2023-06-08	797.43
26	Piura	Trujillo	2022-02-28	2023-02-07	360.84
99	Lima	Piura	2022-03-10	2022-08-06	680.95
37	Cusco	Arequipa	2022-08-04	2023-02-05	145.89
100	Trujillo	Trujillo	2022-02-15	2023-12-17	456.54
43	Cusco	Trujillo	2022-03-10	2022-07-02	194.23
67	Cusco	Cusco	2022-03-29	2023-06-15	431.59
70	Arequipa	Lima	2022-02-18	2022-04-23	645.28

Nos piden crear una variable para la duración del envío, en este caso, en meses.

```
base4 <- base3 %>% mutate(
    duración_envío_meses = trunc(as.integer(difftime(fecha_entrega,fecha_envío,units = "days")/30))
)
head(base4,10) %>%
kable(booktabs = TRUE,format = "latex") %>%
  kable_styling(latex_options = c("striped", "scale_down", "HOLD_position"))
```

id_envío	origen	destino	fecha_envío	fecha_entrega	monto_envío	duración_envío_meses
45	Arequipa	Trujillo	2022-10-11	2023-11-28	62.21	13
12	Lima	Trujillo	2022-01-21	2022-07-09	847.67	5
42	Piura	Trujillo	2022-12-19	2023-06-08	797.43	5
26	Piura	Trujillo	2022-02-28	2023-02-07	360.84	11
99	Lima	Piura	2022-03-10	2022-08-06	680.95	4
37	Cusco	Arequipa	2022-08-04	2023-02-05	145.89	6
100	Trujillo	Trujillo	2022-02-15	2023-12-17	456.54	22
43	Cusco	Trujillo	2022-03-10	2022-07-02	194.23	3
67	Cusco	Cusco	2022-03-29	2023-06-15	431.59	14
70	Arequipa	Lima	2022-02-18	2022-04-23	645.28	2

Agrupamos por origen y destino (región), y obtenemos la duración de envío promedio en meses.

```
prom_envi <- base4 %>% filter(!is.na(duración_envío_meses)) %>%
  select(origen,destino,duración_envío_meses) %>%
  group_by(origen,destino) %>%
  summarise(mean_duración_envío_meses = trunc(mean(duración_envío_meses)))
prom_envi %>%
kable(booktabs = TRUE,format = "latex") %>%
  kable_styling(latex_options = c("striped", "scale_down", "HOLD_position"))
```

origen	destino	mean_duración_envío_meses
Arequipa	Arequipa	6
Arequipa	Cusco	8
Arequipa	Lima	7
Arequipa	No identificado	4
Arequipa	Piura	11
Arequipa	Trujillo	8
Cusco	Arequipa	5
Cusco	Cusco	7
Cusco	Lima	9
Cusco	No identificado	19
Cusco	Piura	8
Cusco	Trujillo	10
Lima	Arequipa	4
Lima	Cusco	3
Lima	Lima	14
Lima	No identificado	8
Lima	Piura	6
Lima	Trujillo	10
No identificado	Arequipa	8
No identificado	Cusco	5
No identificado	Lima	11
No identificado	No identificado	7
No identificado	Piura	5
No identificado	Trujillo	3
Piura	Arequipa	9
Piura	Cusco	7
Piura	Lima	8
Piura	No identificado	4
Piura	Piura	9
Piura	Trujillo	7
Trujillo	Arequipa	5
Trujillo	Cusco	6
Trujillo	Lima	8
Trujillo	No identificado	13
Trujillo	Piura	5
Trujillo	Trujillo	9

Características Dependientes del sistema

Disponibilidad

Medida en que los datos están disponibles y accesibles cuando se necesitan.

https://github.com/bastiannseef/E2422/tree/main/Caso2/data_procesada_caso2

Portabilidad

Para trasladar los datos evaluados bajo el criterio de calidad de la norma ISO/IEC 25012 entre sistemas, guardamos las bases de datos procesadas en archivos ".csv" y también, en formato ".rds" para garantizar eficiencia, portabilidad y seguridad.

Encontrará los siguientes archivos:

- data_caso2.csv: Data general evaluada por los criterios de calidad en formato ".csv".
- data_caso2.rds: Data general evaluada por los criterios de calidad en formato ".rds".
- quantiles_cliente.csv: Data duración promedio de envío en formato ".csv".
- quantiles_cliente.rds: Data duración promedio de envío en formato ".rds".

```
data.frame(base4) %>%
  saveRDS("C:/Users/Sebastian/OneDrive/Escritorio/Repaso/Caso2/data_procesada_caso2/datos_caso2.rds")

readRDS("C:/Users/Sebastian/OneDrive/Escritorio/Repaso/Caso2/data_procesada_caso2/datos_caso2.rds") %>%
  export("C:/Users/Sebastian/OneDrive/Escritorio/Repaso/Caso2/data_procesada_caso2/data_caso2.csv")

data.frame(prom_envi) %>%
  saveRDS("C:/Users/Sebastian/OneDrive/Escritorio/Repaso/Caso2/data_procesada_caso2/prom_envi.rds")

readRDS("C:/Users/Sebastian/OneDrive/Escritorio/Repaso/Caso2/data_procesada_caso2/prom_envi.rds") %>%
  export("C:/Users/Sebastian/OneDrive/Escritorio/Repaso/Caso2/data_procesada_caso2/prom_envi.csv")
```