“janitor” Package in R

Meadow Monticello
Nick Zywalewski

Presentation Outline

1. Package overview
2. Our data set
3. Implementation in R
4. Takeaways and our experience with data cleaning

OoOOoOo! What's that?

What is the “janitor” package?

```
clean_names()
tabyl()
get_dupes()
get_one_to_one()
```

- Published on 12-22 (Meadow’s birthday) -2024 and created by Sam Firke
- Cleaning
  - Parses, appends, converts symbols, and (adds/removes) spaces
  - Removing of empty rows/columns and columns with constant values
  - Converting date formats
- Exploring
  - Search records for duplicates and specific value combinations
  - Inspect one-to-one relationships
  - Building tables
  - Count factor levels


```
# install.packages("janitor") # if not already installed
library(janitor)
```

The Data Set

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	First Name	Last-Name	E-mail(Address)	Secondary Email	Customer ID	Order ID	Order Date (Excel Serial)	Ship Mode	Region	Product Name	Qty	Unit Price ($)	Discount %	Payment Method	Satisfaction Level
2	Joe	Smith	joe.smith17@gmail.com	NA	101	1234	45292	Second Class	East	Stapler - Black	2	5.99	0	Card	Extremely Satisfied
3	Joe	Smith	joe.smith17@gmail.com	NA	101	4567	45382	Second Class	East	Paper Ream	1	4.25	0.1	Card	Extremely Satisfied
4	Emily	Sanders	NA	NA	102	3451	45472	Standard Class	West	Notebook 200p	3	2.5	0	Card	Extremely Dissatisfied
5	NA	Jones	djones12@yahoo.com	NA	NA	5367	45562	Standard Class	South	Binder 3in	1	6.49	0.15	Card	Neutral
6	Dan	NA	dan1976@comcast.net	NA	103	9825	45652	First Class	South	Markers (Set)	1	8.99	0	Card	Satisfied
7	Ava	Miller	avamiller@icloud.com	NA	104	7548	45742	Standard Class	West	Paper Ream	5	4.25	0.05	Card	Satisfied
8	Ava	Miller	avamiller@icloud.com	NA	104	1679	45832	Standard Class	West	Stapler - Black	2	5.99	0	Card	Dissatisfied
9	Michael	Nelson	mikenelson@gmail.com	NA	105	4875	45922	First Class	Central	Desk Chair	1	89.99	0.2	Card	Extremely Satisfied
10	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	Hank	Johnson	hank_j12@gmail.com	NA	106	1683	45931	Second Class	East	Desk Lamp	1	24.5	0	Card	Satisfied

```
# Loading data
orders_raw <- read.csv("C:/Users/nzywa/Downloads/DSS445/Datasets/messy_retail_orders.csv", header=TRUE)
orders_raw
```

Implementation of Package with Data Set

janitor package be like...

Cleaning: Column Names

```
names(orders_raw) # Look at messy column names

orders <- clean_names(orders_raw) # Clean column names
names(orders) # Look at new column names
```

This cleans the column names of our dataset, ensuring that there are no funky symbols or spaces in them.

Notice how the output will utilize underscores to fix this.

Cleaning: Dropping Empty Rows and Columns

```r
dim(orders) # Check dimension of dataset prior to dropping anything

# Dropping empty rows and columns
orders <- remove_empty(orders, which = c("rows", "cols"))
dim(orders) # checking the dimension again to see changes

names(orders) # checking which column(s) was/were dropped
```

Here, we remove empty rows and columns.

Notice how there is one row and one column that get dropped after this cleaning step.

10X15 → 9X14

Cleaning: Removing Columns with Constant Values

```
# dropping columns with a constant value
orders <- remove_constant(orders)
names(orders) # inspecting which column(s) was/were dropped
```

Since everyone paid with a credit card, notice how “Payment Method” is no longer in our list of column names.

Exploring: Duplicates

```
# Finding which customers appear in the dataset more than once
dupes <- get_dupes(orders, customer_id)
dupes
```

Here, we are taking a look at rows that have similar data in them. In this particular example. We are examining which customers placed multiple orders.

Notice how the output allows us to see all of the details in these cases.

Exploring: Table Creation

```r
# Creating a table to summarize what products were purchased
product_table <- tabyl(orders, product_name)
product_table
```

```r
# Two way table for shipment mode and region
tab_ship_region <- tabyl(orders, ship_mode, region)
tab_ship_region
```

Here, we are building tables consistent with what variable(s) we are interested in. One-way and two-way tables help us see relationships between variables and to quickly count frequencies, compare groups, and identify patterns in the data.

Our Experience and Takeaways

- DSS 416, Data Wrangling and Visualization
  - Fe Y Alegría
- DSS 420, Data Mining
- DSC 225, Data Science of Sports
- MAT 470, Statistics in Research

These classes all dealt with filthy data. Utilizing “janitor” would have helped tremendously. I am curious if there exists similar packages in Python and if so how they might differ.

References

- https://cran.r-project.org/web/packages/janitor/index.html
- https://cran.r-project.org/web/packages/janitor/vignettes/janitor.html#clean-dataframe-names-with-clean_names
- https://www.rdocumentation.org/packages/janitor/versions/2.2.1
- https://www.r-bloggers.com/2024/08/top-25-r-packages-you-need-to-learn-in-2024/

Also the “help” function came in handy!

Any questions? Thank you!

Now scram!!!

Oscar the Grouch when the data isn't clean. JK he loves the grime.