# Instacart Capstone Project - Data Wrangling

Since the dataset was obtained from Kaggle, the data is already pretty clean. The only missing values are inside the days_since_prior_order column of orders.csv. There are some "NaN" values in that column which most likely represent the first order of a particular user. Since it's the first order, there is no prior order so days_since_prior_order is "NaN".

```
print(orders.info())
orders.head()
```
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3421083 entries, 0 to 3421082
Data columns (total 7 columns):
order_id                 int64
user_id                  int64
eval_set                 object
order_number             int64
order_dow                int64
order_hour_of_day        int64
days_since_prior_order   float64
dtypes: float64(1), int64(5), object(1)
memory usage: 182.7+ MB
None
```

|   | order_id | user_id | eval_set | order_number | order_dow | order_hour_of_day | days_since_prior_order |
|---|----------|---------|----------|--------------|-----------|-------------------|------------------------|
| 0 | 2539329  | 1       | prior    | 1            | 2         | 8                 | NaN                    |
| 1 | 2398795  | 1       | prior    | 2            | 3         | 7                 | 15.0                   |
| 2 | 473747   | 1       | prior    | 3            | 3         | 12                | 21.0                   |
| 3 | 2254736  | 1       | prior    | 4            | 4         | 7                 | 29.0                   |
| 4 | 431534   | 1       | prior    | 5            | 4         | 15                | 28.0                   |

The "NaN" comprises about 6% of total number of rows in the orders data. When doing the EDA involving the "days_since_prior_order" column, these "NaN" data are excluded. In the modeling part of the project, these "NaN" can be replaced by "-1" and the model should be able to learn what is the meaning of "-1".