Data analysis and presenting results

Syntax

The groupby ('column name') method for grouping data

In
```
df.groupby('column name')

# group by column_1
# and output column_2
df.groupby('column_1')['column_2']

# count the quantity in a group
df.groupby('column name').count()

# calculate the total in the group
df.groupby('column name').sum()
```

The sort_values (by = 'column name') method for sorting a table by the selected column

In
```
# sort in ascending order
# (by default)
df.sort_values(by = 'column name')

# sort by descending order
df.sort_values(by = 'column name',
    ascending = False)
```

The max ( ) method for identifying the maximum value

In
```
df['column name'].max()
# maximum value in the column
```

The min ( ) method for identifying minimum value

In
```
df['column name'].min()
# minimum value in the column
```

The mean ( ) method for calculating the arithmetic mean

In
```
df['column name'].mean()
# mean value of the column
```

The median ( ) method for calculating the median

In
```
df['column name'].median()
# the median of the column
```

Glossary

Grouping
Splitting data into groups based on a certain feature

We can describe the grouping stages with this formula:
split-apply-combine

- *split* into groups based on a particular criterion
- *apply* a method to each group individually, for example, calculating the size of a group with the count() method or summing up by calling sum()
- *combine* results into a new data structure. Depending on the separation conditions and the method execution, that could be DataFrame and Series.

The logical operator **&** (similar to **AND**)
serves to join multiple conditions into a single one with logical indexing

Minimum and **maximum**
The largest and smallest numbers in a set.
The maximum and minimum values are usually calculated based on separate features

Mean and **Median**
Used to estimate the values at the middle of the data set. If the data set is uniform without many outlying values, the mean works. But when there are values that significantly diverge from the main set, they can pull the mean with them. In that case, we use the median.

Well-ordered results can help get across the main idea and value behind the research.
A few pieces of advice:

- When showing how a parameter changes over time, put its values in a row and define the time ranges in the columns
- If you need to show different attributes for competing categories (such as genres), put each category on a separate row, with attribute values in the columns
- You don't have to put all the data in one table. It can be better to have multiple tables so that each one can reflect one important idea
- Detailing from largest to smallest works beautifully. Have more detailed tables to go with your main summary table. Here's an example. First you have an overview table that covers some entire service, then you have more detailed tables, for summaries by user group, city, etc.