

Data wrangling steps

3 questions I have to answer:

1. **What kind of cleaning steps did you perform? (see below)**
2. **How did you deal with missing values, if any?**
3. **Were there outliers, and how did you handle them?**

Data Wrangling

CBG patterns.csv

CBG geographic data.csv

STEP 1

Import csv CBG patterns.csv

Observe head

Observe data shape

Observe type

Delete unnecessary columns

Check if null rows, remove null rows if necessary in key column

Observe type of columns

Fix column type if necessary

census_block_group	float to object (done in import)
visitor_home_cbgs	object to float (imported as dictionary, no need to convert)
visitor_work_cbgs	object to float (imported as dictionary, no need to convert)
popularity_by_hour	object to float (imported as list, no ne ed to convert)
popularity_by_day	object to float (imported as dictionary, no need to convert)

Check statistics

Import csv CBG geographic data.csv

Observe head

Observe data shape

Observe type

Delete unnecessary columns

Check if null rows, remove null rows if necessary in key column

Observe type of columns

Fix column type if necessary

Check statistics

Find

- rows that are present in data1 and in data2
- rows that are present in data1 and not in data2
- rows that are present in data2 and not in data1

STEP 2

Identify and group columns

2.1 Key column

Select only column 'census_block_group'

→ **Check consistency of data**

Check first charac is a nb

Check last charact is a nb

Check length

Check that column 'census_block_group' is a unique identifier for both datasets

2.2 Columns that start with {

Select only columns 'visitor_home_cbgs', 'visitor_work_cbgs', and 'popularity_by_day'

→ **Prepare data for further checks**

Add index
Unpivot other columns than Index
Remove {
Remove }
Split column by delimiter ','
Unpivot other than 'variable'
Calculate position of delimiter ':'

→ **Check consistency of data**

Filter on 'visitor_home_cbgs' + 'visitor_work_cbgs'
Group by on position of ':' → check that the delimiter is always in same position and *null*
Filter 'popularity_by_day'
Split column by delimiter ':'
Group by on position of ':' → check that there is a all days of week in each row, order is the same

2.3 Columns of type float (numeric)

Select only columns 'raw_visit_count', 'raw_visitor_count', and 'distance_from_home'
Add index
Unpivot other columns than Index
Check that all lines are number

2.4 Columns containing text of brands

Select only columns 'related_same_day_brand', 'related_same_month_brand', and 'top_brands'

2.5 Column 'popularity by hour'

Select only columns 'popularity_by_hour'

STEP 3

Split columns 'visitor_home_cbgs' (vis_home), 'visitor_work_cbgs' (vis_work), 'popularity_by_day' (pop_day) and 'popularity_by_hour' (pop_h),
Once split is done, check for NaN
Remove the rows that are full of NaN
Append 'pop_day'/'vis_home'/'vis_work' to "common2"
Convert new columns to numeric (float)

STEP 4

Calculate ratio for week and day

STEP 5

Actions of columns 'brands'

- calculate the top 10 of brands across all cbg for each three brand-related columns
- use the variable 'overall_top10' created above, and create a column for each existing brand, and assign a 0 or a 1 depending if it is a top 10 brand for the cbg (binary)