Knowledge Discovery and Data Mining

Prof. Dr. SCHOMMER Christoph

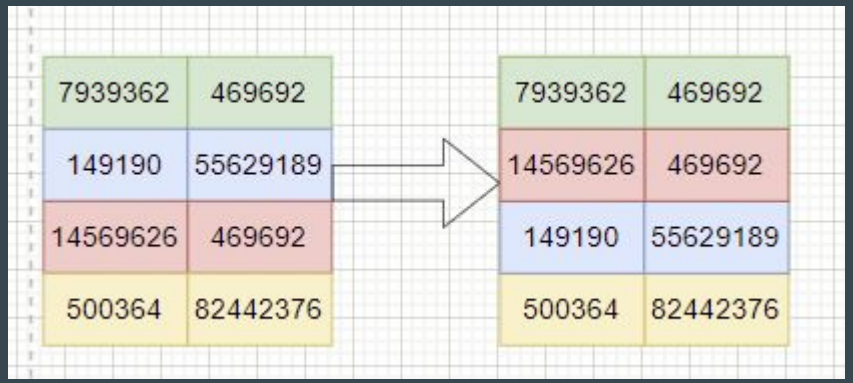
•••

Université du Luxembourg

Diabetes data set

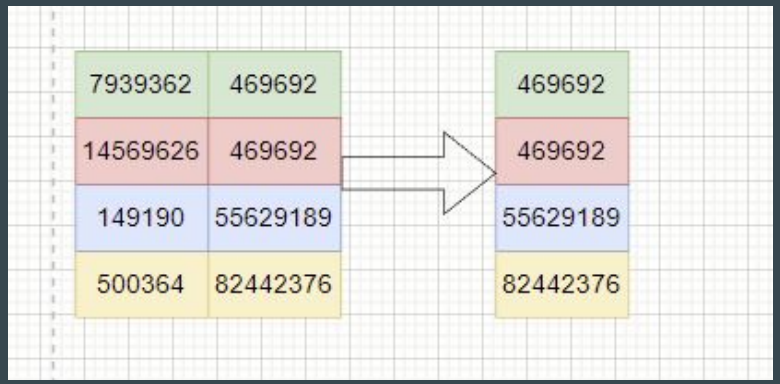
Guillaume BALLINGER - Data cleaning
Filipe DA SILVA - Analyse
Esada LICINA - Visualisation

Data preprocessing (Multi-level grouping)



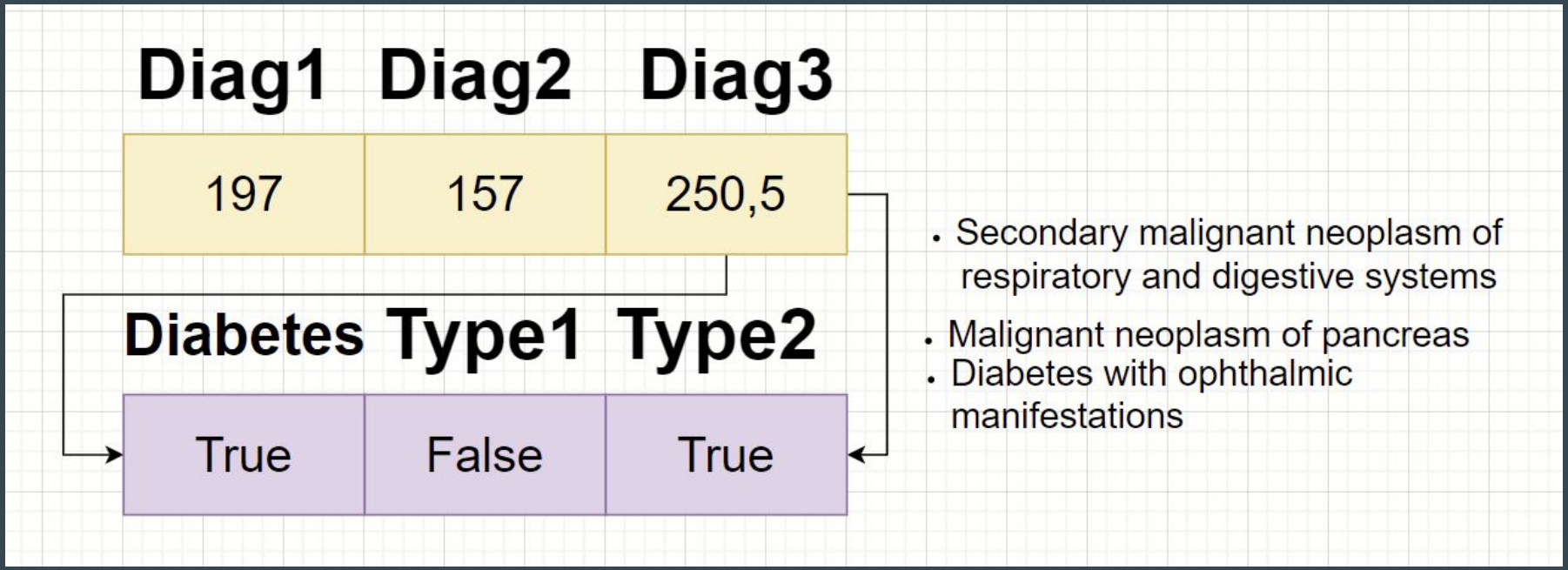
- Encounter ID
- Patient ID

Data preprocessing (Drop column)

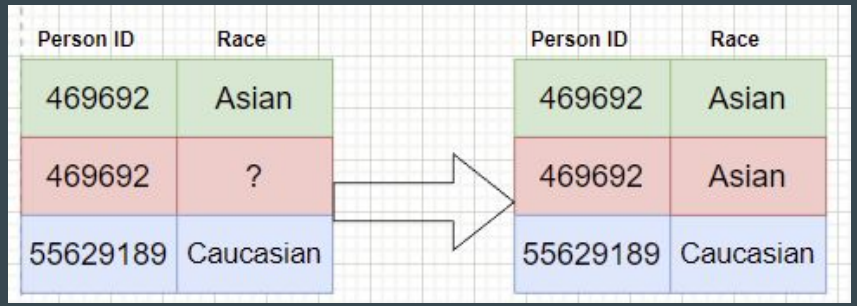


- Payer code
- Encounter ID
- Diagnoses
- Citoglipton

Data preprocessing (Diagnosis)



Data preprocessing (Drop Imputation)

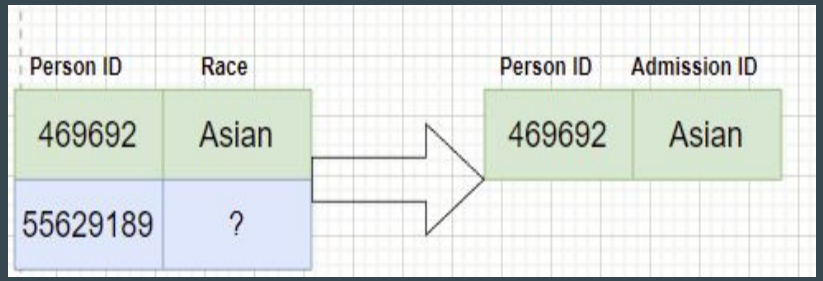


Person ID	Race
469692	Asian
469692	?
55629189	Caucasian

Person ID	Race
469692	Asian
469692	Asian
55629189	Caucasian

- Deterministic
- Pediatrics
- Probabilistic?
  - No ⇒ global average

Data preprocessing (Listwise deletion)



⚠️⚠️Last resort ⚠️⚠️

Data preprocessing (General)

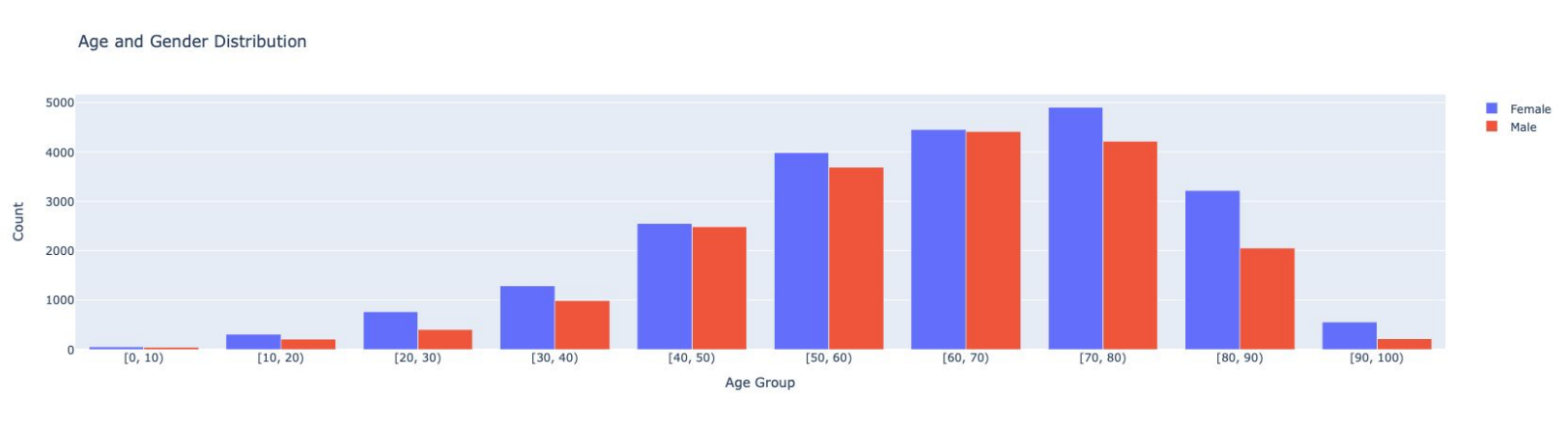
101763 rows only 55440 cases had diabetes

18k rows have been repaired

101763 rows ⇒ 40782 rows

Diabetes patient records for the analyses

It shows that female are a bit more affected starting by the age of 20



Data Precision, recall and accuracy

Diabetes type 1 - Recall looks very low with 12% - it is linked to True Negative

Classification	Positive diagnose	Negative diagnose	Precision	68%
Positive diagnose	2663	70	Recall	12%
Negative diagnose	1243	20385	Accuracy	95%

Data Precision, recall and accuracy

Diabetes type 2

Classification	Positive diagnose	Negative diagnose	Precision	97%
Positive diagnose	14835	3213	Recall	82%
Negative diagnose	385	3369	Accuracy	83%

Whey the difference in the recall for type 1 diabetes

It is more likely that you have a type 2 diabetes vs. type 1 diabetes

Check given that out of 40 782 test, if you take out duplicates.

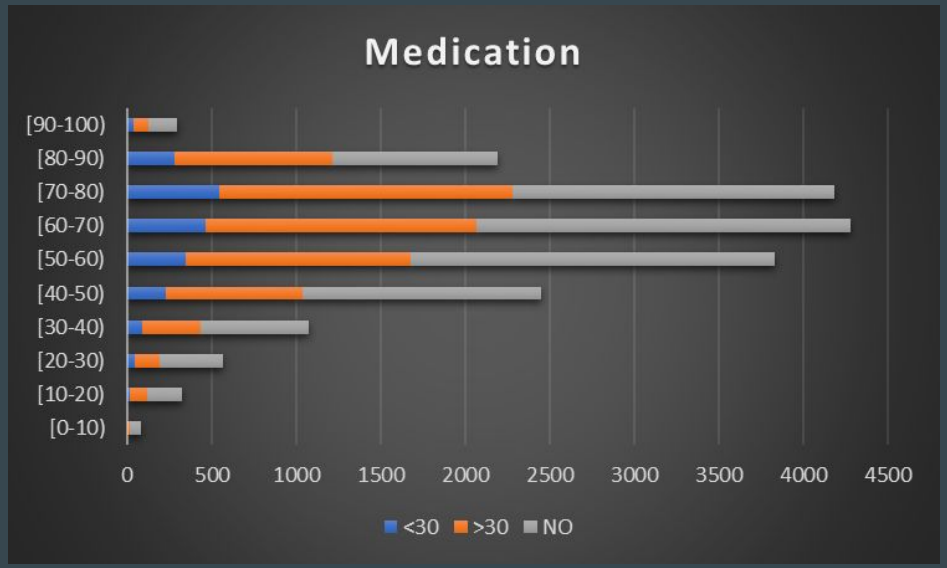
Diabetes type 2 has 14835 positive results.

Diabetes type 1 has 2663 positive results.

Type 2 makes 85% of the positive results.

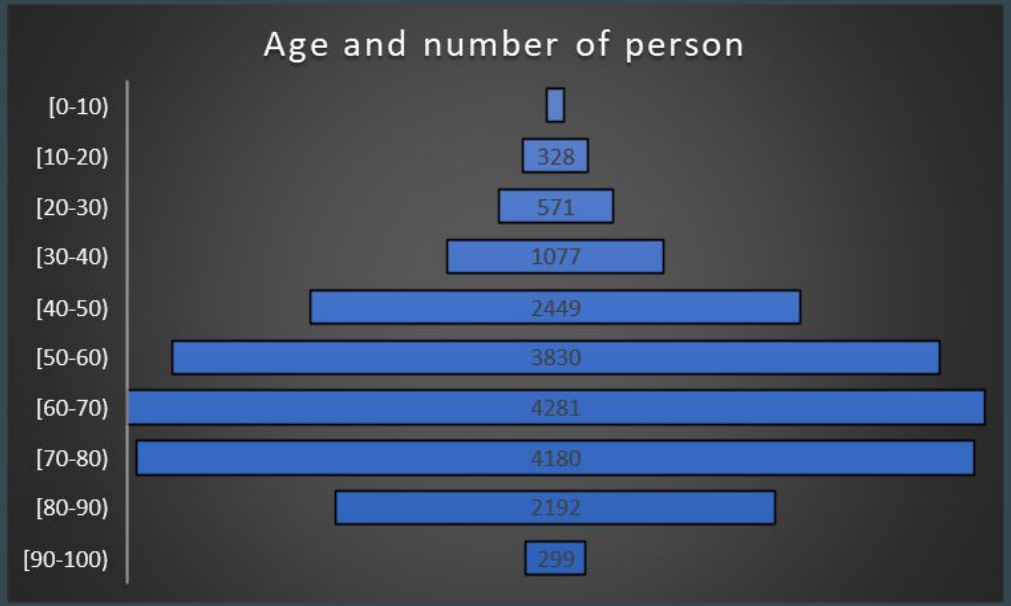
Medication

If you take medication, has a very clear impact in the number of visits in the hospital



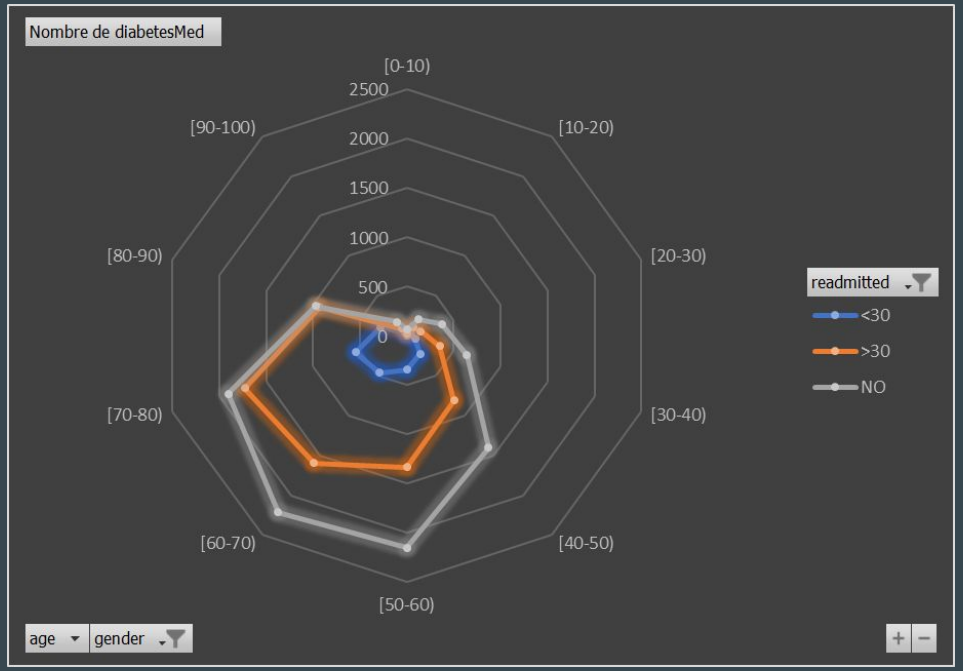
Age

Diabetes increase with age as the number of patient is linked with age



Interpret and discuss the results - Medication and age

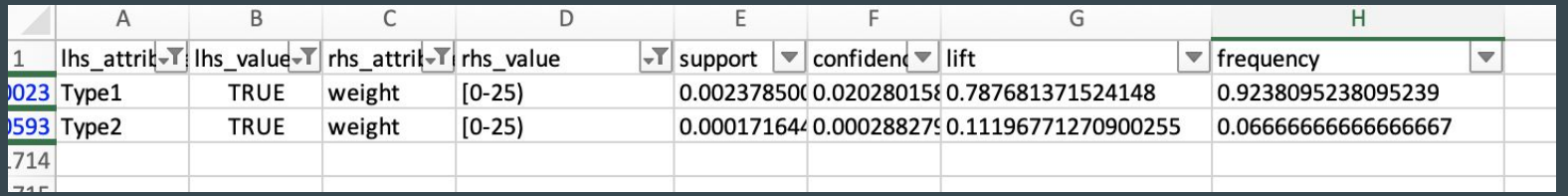
Starting at age of 50 diabetes medication is increase and at as age is augmenting the number of hospitalization is increasing.



Checking for association

	A	B	C	D	E	F	G	H	I
1	lhs_attrib	lhs_value	rhs_attrib	rhs_value	support	confidenc	lift	frequency	
70020	Type1	TRUE	weight	[50-75)	0.084375459	0.719422956	0.8835071971805709	0.10361960973259456	
70021	Type1	TRUE	weight	[25-50)	0.005737825	0.048923269	0.6520224815830570	0.7647058823529411	
70022	Type1	TRUE	weight	[75-100)	0.024790348	0.211373614	0.91034397266903600	0.1411419796174787	
70023	Type1	TRUE	weight	[0-25)	0.002378500	0.020280158	0.787681371524148	0.9238095238095239	
70590	Type2	TRUE	weight	[50-75)	0.481266244	0.808294209	0.9926479902461423	0.5910322813779813	
70591	Type2	TRUE	weight	[25-50)	0.001152469	0.001935590	0.25796482830015294	0.15359477124183007	
70592	Type2	TRUE	weight	[75-100)	0.112819381	0.189481920	0.788010180946200	0.6423286332542231	
70593	Type2	TRUE	weight	[0-25)	0.000171644	0.000288275	0.11196771270900255	0.06666666666666667	
71714									
71715									

Checking diff in type 1 vs type 2 diabetes

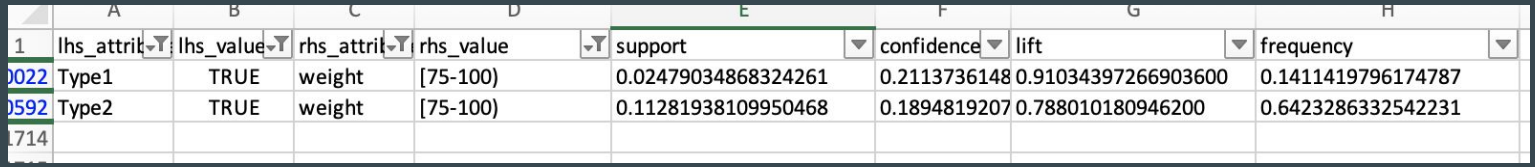


	A	B	C	D	E	F	G	H
1	lhs_attrib	lhs_value	rhs_attrib	rhs_value	support	confidenc	lift	frequency
0023	Type1	TRUE	weight	[0-25)	0.00237850(	0.020280158	0.787681371524148	0.9238095238095239
0593	Type2	TRUE	weight	[0-25)	0.000171644	0.000288279	0.11196771270900255	0.06666666666666667
1714								

To check for children as weight is from 0 to 25 kg, it is indicate that for type 1 has a greater frequency of 0.92 vs type 2 of 0.06. The lift is as well greater in type 1 for 0.78 vs type 2 of lift 0.11.

Type 1 diabetes is more frequent for children.

Checking diff in type 1 vs type 2 diabetes



	A	B	C	D	E	F	G	H
1	lhs_attrib	lhs_value	rhs_attrib	rhs_value	support	confidence	lift	frequency
0022	Type1	TRUE	weight	[75-100)	0.02479034868324261	0.2113736148	0.91034397266903600	0.1411419796174787
0592	Type2	TRUE	weight	[75-100)	0.11281938109950468	0.1894819207	0.788010180946200	0.6423286332542231
1714								

To check for adults as weight is from 75 to 100 kg, it is indicate that for type 1 has a lower frequency of 0.14 vs type 2 of 0.64. The lift is as well greater in type 1 for 0.91 vs type 2 of lift 0.78.

It show that person with a high level of weight are likely to have diabetes

Difference frequency and lift

The frequency and lift are both measures of association between two variables, and in this case, they indicate that there is a stronger association between the type of diabetes and the given weight range.

The frequency indicates the proportion of observations belonging to a particular category, while lift measures the strength of association between two variables.

Association table

Challenging part and self-included

DataTable function

DataTable function

DataTable function

DataTable function

Assosiation Table

Show 10 entries

Chose left attribute:	Filter lhs Values:	Chose right attribute:	Filter rhs Values:	Filter Support:	Filter Confidence:	Filter Lift:	Filter Frequency:
None		None					
A1Cresult	Norm	admission_type_id	2	0.009710166249816	0.2163934426229508	1.1598051487776555	0.0520436325404126
A1Cresult	>7	num_medications	15	0.0015202785542641	0.0451565914056809	0.8232347388048646	0.02771569065713
A1Cresult	>7	num_medications	16	0.0018880878819086	0.056081573197378	1.0662558126505688	0.0358974358974358
A1Cresult	>7	num_medications	17	0.0015202785542641	0.0451565914056809	0.987440273837256	0.0332439678284182
A1Cresult	>7	num_medications	18	0.0018635672600657	0.0553532410779315	1.326331302961342	0.044653349001175
A1Cresult	>7	num_medications	19	0.0012505517139914	0.0371449380917698	0.9569455876554376	0.0322173089071383
A1Cresult	>7	num_medications	20	0.0011769898484625	0.0349599417334304	1.0033331061032795	0.0337790288529204
A1Cresult	>7	num_medications	21	0.0011524692266195	0.0342316096139839	1.1556568735740849	0.0389072847682119
A1Cresult	>7	num_medications	22	0.0010298661174047	0.0305899490167516	1.1702807699823317	0.0393996247654784
A1Cresult	>7	num_medications	23	0.0006130155460742	0.0182083029861616	0.8428728857907448	0.0283768444948921

Showing 1 to 10 of 31,606 entries (filtered from 71,712 total entries)

Previous 1 2 3 4 5 ... 3161 Next

Selecting Method

Python Code

```
with open("association2NoFilter.csv", "r") as f:
    df = pd.read_csv(f)
column_names = list(df.columns)

unique_values = {}
for column in column_names:
    unique_values[column] = list(df[column].unique())
first = df["lhs_attribute"].unique()
second = df["rhs_attribute"].unique()
first = np.insert(np.asarray(first), 0, "None")
second = np.insert(np.asarray(second), 0, "None")

return render_template('server_table.html', title='Assosiation Table', l_column_names=first, r_column_names=second)
```

```
if "latt" in req and req["latt"] is not None:
    leftAttribute = req["latt"]
if "ratt" in req and req["ratt"] is not None:
    rigthAttribute = req["ratt"]

if leftAttribute != "None":
    data_filter = data_filter.query('lhs_attribute == "{}"'.format(leftAttribute))
else:
    data_filter = data_filter
if rigthAttribute != "None":
    data_filter = data_filter.query('rhs_attribute == "{}"'.format(rigthAttribute))
else:
    data_filter = data_filter
```

Selecting Method

HTML with Javascript

```
<th>
    <label for="latt">Chose left attribute:</label>
    <form id="my-form1" action="{{ url_for('api_data') }}" method="post">
        <select id="latt" name="latt">
            <option value="None">None</option>
            {% for column in l_column_names %}
            <option value="{{ column }}" {% if column == l_column_name %}selected{% endif %}>{{ column }}</option>
            {% endfor %}
        </select>
    </form>
</th>
```

```
$(document).ready(function() {

    // Listen for the change event of the select element
    $('#latt').on('change', function() {
        // Submit the form when the value changes
        $('#my-form').submit();
    });
```

```
$('#my-form1').on('change', function(event) {
    event.preventDefault();
    var formData = $(this).serialize();
    $.ajax({
        url: '/api/data',
        data: formData,
        dataType: 'json',
        type: 'GET',
        success: function(data) {
            table.clear();
            table.rows.add(data).draw();
        },
        error: function(xhr, ajaxOptions, thrownError) {
            console.log(thrownError);
        }
    });
});
```

Searching Method

Python Code

```
search_support_ = request.args.get('support_search')
```

```
if search_support_ is not None or search_support != "":
    if search_support_ != "":
        try:
            data_filter = data_filter.query('support {}'.format(search_support_))
            search_support = search_support_
        except:
            try:
                data_filter = data_filter.query('support {}'.format(search_support))
            except:
                pass
    else:
        data_filter = data_filter
        search_support = ""
```

Searching Method

HTML with Javascript

```
<th>
    <label for="support_search">Filter Support:</label>
    <input type="text" id="support_search" name="support_search" class="form-control form-control-sm">
</th>
```

```
 $('#support_search').on('keyup', function () {
  table.column('support:name').search($('#support_search').val()).draw()
});
```

Questions?