## Data Preprocessing

1. Use psycopg2 to connect Pgadmin.
2. Execute queries to get age, special_measure, gender, mobility, and if resolved
3. Plot the data to feel the mobility data as well as the age group.
4. Checked no null data since all data is preprocessed in Phase 2.
5. Remove data that has UNKNOWN gender
6. Turn Age group, month, gender to numeric data as Figure 1.
7. Turn special measure to one hot encoding
8. Now the data looks like Figure 2.
9. Normalize the data between 0 to 1. The data looks like Figure 3.

Figure 1:

```
mapping_age = {
    '<20':0,
    '20s':1,
    '30s':2,
    '40s':3,
    '50s':4,
    '60s':5,
    '70s':6,
    '80s':7,
    '90+':8
}
```

```
mapping_month = {
        11:1,
        12:2,
        1:3,
        2:4
    }
```

```
mapping_gender = {
    'MALE':0,
    'FEMALE':1,
    'UNSPECIFIED':2,
    'GENDER DIVERSE':3
}
```

Figure 2:

| month | Age_Group | gender | mobility | resolved | special_measure_Control | special_measure_Lockdown | special_measure_Prevent | special_measure_Protect | special_m |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 0 | -26.0 | 1 | 0 | 0 | 0 | 0 | |
| 1 | 2 | 0 | -26.0 | 1 | 0 | 0 | 0 | 0 | |
| 1 | 4 | 0 | -26.0 | 1 | 0 | 0 | 0 | 0 | |
| 1 | 0 | 1 | -26.0 | 1 | 0 | 0 | 0 | 0 | |
| 1 | 2 | 1 | -26.0 | 1 | 0 | 0 | 0 | 0 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 4 | 0 | 1 | -23.0 | 0 | 1 | 0 | 0 | 0 | |
| 4 | 1 | 0 | -23.0 | 0 | 1 | 0 | 0 | 0 | |
| 4 | 1 | 1 | -23.0 | 0 | 1 | 0 | 0 | 0 | |
| 4 | 4 | 1 | -23.0 | 1 | 1 | 0 | 0 | 0 | |
| 4 | 2 | 0 | -23.0 | 0 | 1 | 0 | 0 | 0 | |

Figure 3:

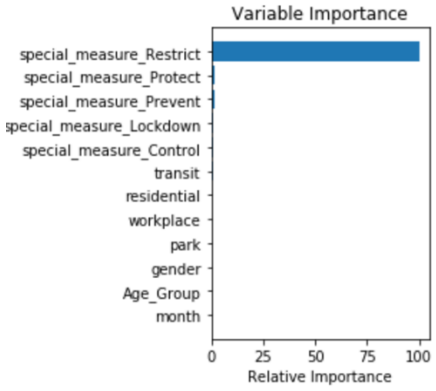
| month | Age_Group | gender | mobility | resolved | special_measure_Control | special_measure_Lockdown | special_measure_Prevent | special_measure_Protec |
|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.375 | 0 | 0.084112 | 1 | 0 | 0 | 0 | 0 |
| 0.0 | 0.250 | 0 | 0.084112 | 1 | 0 | 0 | 0 | 0 |
| 0.0 | 0.500 | 0 | 0.084112 | 1 | 0 | 0 | 0 | 0 |
| 0.0 | 0.000 | 1 | 0.084112 | 1 | 0 | 0 | 0 | 0 |
| 0.0 | 0.250 | 1 | 0.084112 | 1 | 0 | 0 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | .. |
| 1.0 | 0.000 | 1 | 0.093458 | 0 | 1 | 0 | 0 | 0 |
| 1.0 | 0.125 | 0 | 0.093458 | 0 | 1 | 0 | 0 | 0 |
| 1.0 | 0.125 | 1 | 0.093458 | 0 | 1 | 0 | 0 | 0 |
| 1.0 | 0.500 | 1 | 0.093458 | 1 | 1 | 0 | 0 | 0 |
| 1.0 | 0.250 | 0 | 0.093458 | 0 | 1 | 0 | 0 | 0 |

## Obtained from Data

Since the data with ‘fatal’ label only contain a small part of total, which could lead to unbalanced. In this case, the classifiers do not act well, although the accuracy is high, the recall is almost zero. Thus, we first balanced the data, select the same number of ‘un-fatal’ data as ‘fatal’ data.

From the ‘decision tree’ classifier and ‘Random Forest’ classifier, it obviously shows the ‘Age’ feature has a high correlation with the ‘fatal’. From the ‘Random Forest’ mainly factors (shows belowing, left), we can see that ‘Age’ has 0.79452 correlation with ‘fatal’, which is much higher than the second factor. From the decision of the tree, we can also see, ‘Age’ is the decisive factors in many split points. Moreover, it shows that age is larger, the probability of ‘fatal’ is higher, which means COVID-19 is more dangerous to the elderly people. Another factors which could lead split is ‘Transit’ and ‘Park’, it shows when more people have outside activity, the probability of ‘fatal’ is higher than normal, since people’s activity exacerbated the spread of the virus.

```
 1) Age_Group                    0.784520
 2) special_measure_Control      0.052139
 3) month                        0.041303
 4) transit                      0.025591
 5) workplace                    0.024810
 6) park                         0.024145
 7) residential                  0.019147
 8) gender                       0.010140
 9) special_measure_Restrict     0.007871
10) special_measure_Prevent      0.005257
11) special_measure_Lockdown     0.004356
12) special_measure_Protect      0.000721
```




From the ‘Gradient Boosting’ classifier, we can see that ‘special_measure_Restrict’ has the high weight in deciding ‘fatal’ label after 15 estimators (shows above, right). It shows the same result as factors ‘Transit’ and ‘Park’, since ‘restrict’ do not have too much limit of people’s activity compare with other measures.

From these three classifiers, we can find that the ‘fatal’ has positive relation with people’s outside activity, as more people outside, the probability of ‘fatal’ is higher overall, and with all the people who are infected with COVID-19, the elderly people have high ‘fatal’ risk.