MIS548 Milestone Report

Group 7: Bo Huang, Jaejun Shin, Jaxson Schaefer, Katie Scott, Zachary Bartlett

We have identified our dataset, Swiss Dwellings, and have started an initial review. Once our end goal is determined, we will clean the data to perform our analysis.

[Swiss Dwellings: A large dataset of apartment models including aggregated geolocation-based simulation results covering viewshed, natural light, traffic noise, centrality and geometric analysis (zenodo.org)]

Swiss Dwellings is a large dataset of apartment models including aggregated geolocation-based simulation results covering viewshed, natural light, traffic noise, centrality, and geometric analysis. The link to the dataset provided four csv files to download. These were then uploaded into Python where we looked at the head, shape, count, description, and datatypes for each csv. After this we combined all into one dataframe for easier comparisons and analysis later.

Question ideas for analysis of Swiss Dwellings

- Analyze location type, climate, and amenities nearby to see if they are related to the overall rating of the property.
- Review dwelling trends over time. Does the most popular dwelling change over time?
- Analyze interactions between various variables.
- Visualize and show raw data/statistics
    - Example: Visualize how many dwelling types there are or how many are near a certain amenity (shop).
    - Create visuals of the data on a map with availabilities shown on the map.

Geometries dataset columns

```
print(geom.columns)
```

```
Index(['apartment_id', 'site_id', 'building_id', 'plan_id', 'floor_id',
       'unit_id', 'area_id', 'unit_usage', 'entity_type', 'entity_subtype',
       'geometry', 'elevation', 'height'],
      dtype='object')
```

Location Ratings dataset columns

```
print(lr.columns)
```

```
Index(['building_id', 'location_rating_MIKRAT_W', 'location_rating_
IMAGE_W',
       'location_rating_FZ_W', 'location_rating_DL_W',
       'location_rating_NASE_W_DOM', 'location_rating_FGFRQZ'],
      dtype='object')
```

Simulations dataset columns (sample)

```
print(sim.columns)
```

```
Index(['site_id', 'building_id', 'plan_id', 'floor_id', 'unit_id',
'area_id',
       'unit_usage', 'apartment_id', 'layout_compactness',
       'layout_is_navigable',
       ...
       'connectivity_balcony_distance_stddev',
       'connectivity_loggia_distance_max', 'connectivity_loggia_dis
tance_mean',
       'connectivity_loggia_distance_median',
       'connectivity_loggia_distance_min', 'connectivity_loggia_dis
tance_p20',
       'connectivity_loggia_distance_p80',
       'connectivity_loggia_distance_stddev',
       'layout_biggest_rectangle_length', 'layout_biggest_rectangle
_width'],
      dtype='object', length=369)
```

Locations dataset columns (sample)

```
print(local.columns)
```

```
Index(['building_id', 'climate_tnorm_year', 'climate_tnorm_januar
y',
       'climate_tnorm_february', 'climate_tnorm_march', 'climate_tn
orm_april',
       'climate_tnorm_may', 'climate_tnorm_june', 'climate_tnorm_ju
ly',
       'climate_tnorm_august',
       ...
       'walkshed_shop_caravan', 'walkshed_shop_water',
       'walkshed_healthcare_veterinary', 'walkshed_shop_swimming_po
ol',
       'walkshed_historic_baptistry', 'walkshed_shop_houseware;elec
tronics',
       'walkshed_shop_pyrotechnics;party', 'walkshed_historic_vehic
le',
       'walkshed_amenity_lavoir', 'walkshed_healthcare_speech_thera
pist'],
      dtype='object', length=503)
```

Then combined the separate data frames into one called Swiss.

Sampling of columns from combined dataset Swiss

```
[5 rows x 892 columns]

 Index(['apartment_id', 'site_id', 'building_id', 'plan_id', 'floor
_id',
       'unit_id', 'area_id', 'unit_usage', 'entity_type', 'entity_s
ubtype',
       ...
       'connectivity_balcony_distance_stddev',
       'connectivity_loggia_distance_max', 'connectivity_loggia_dis
tance_mean',
       'connectivity_loggia_distance_median',
       'connectivity_loggia_distance_min', 'connectivity_loggia_dis
tance_p20',
       'connectivity_loggia_distance_p80',
       'connectivity_loggia_distance_stddev',
       'layout_biggest_rectangle_length', 'layout_biggest_rectangle
_width'],
      dtype='object', length=892)
```

Once the dataset has been combined, we will go through cleansing the dataset. This will involve going through Null values and either eliminating them from the data if it is an insignificant number of data points or looking at averages of quantitative columns. This process will also take standardizing columns with words and letters to ensure punctuation and other items do not interfere with our analysis.