Pipelines (Airflow or Luigi)

- Task Description:
    1. Define preprocessing from previous step as Luigi pipeline/Airflow dag (sequence of tasks)
    2. Each task should be as thin as possible
    3. Types of tasks that could be included:
        - merging with another data source (if can't find it, split existing data in some reasonable manner)
        - cleaning (filtering, correct data types)
        - split (both train/test and cross-validation). Could be done in several ways, choose one reasonable for you dataset.
        - unsupervised transformations (OHE, embeddings)
- Criteria:
    1. Luigi/Airflow best practices
    2. Correct split with respect to data
    3. Clean code / code structure
- Materials:
    1. Airflow:
        a. https://airflow.apache.org/docs/apache-airflow/stable/index.html
        b. https://airflow.apache.org/docs/apache-airflow/stable/concepts.html#jinja-templating
        https://airflow.apache.org/docs/apache-airflow/stable/macros-ref.html
        c. https://youtube.com/playlist?list=PLYizQ5FvN6pvIOcOd6dFZu3lQqc6zBGp2
    2. Luigi:
        a. https://luigi.readthedocs.io/en/stable/
        b. https://www.datarevenue.com/en-blog/how-to-scale-your-machine-learning-pipeline
        c. https://github.com/cbohara/luigi_data_pipeline
        d. https://marcobonzanini.com/2015/10/24/building-data-pipelines-with-python-and-luigi/
        e. https://github.com/jondinu/data-engineering-101
        f. Overview with configs and dates:
        g. http://bytepawn.com/luigi.html
        h. yield in Luigi:
        i. https://vsupalov.com/luigi-multiple-requires/
        j. https://towardsdatascience.com/building-spotify-discover-weekly-email-alert-with-luigi-ca0bc800d137
        k. https://datahovel.com/2016/07/19/how-to-create-a-data-pipeline-using-luigi/