# **ETL Project Report:**

**Database:** Top performing high schools in Applecross (Schools\_DB)

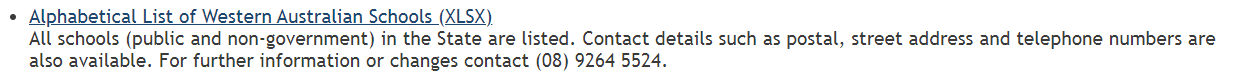
**Perceived Customer:** A family has moved into Applecross and is seeking to enrol their Year 12 child in the best performing government secondary school in the Applecross area based on its performance during the 2020 ATAR exam, and would also like to consider how many children are enrolled in Year 12 to ensure their child gets adequate attention.

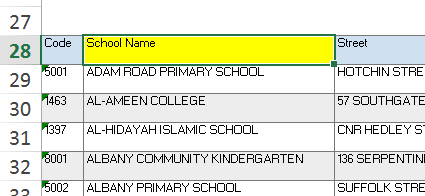
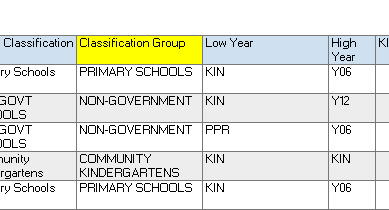
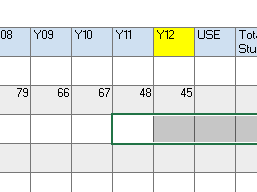
## **Project Timeline:**

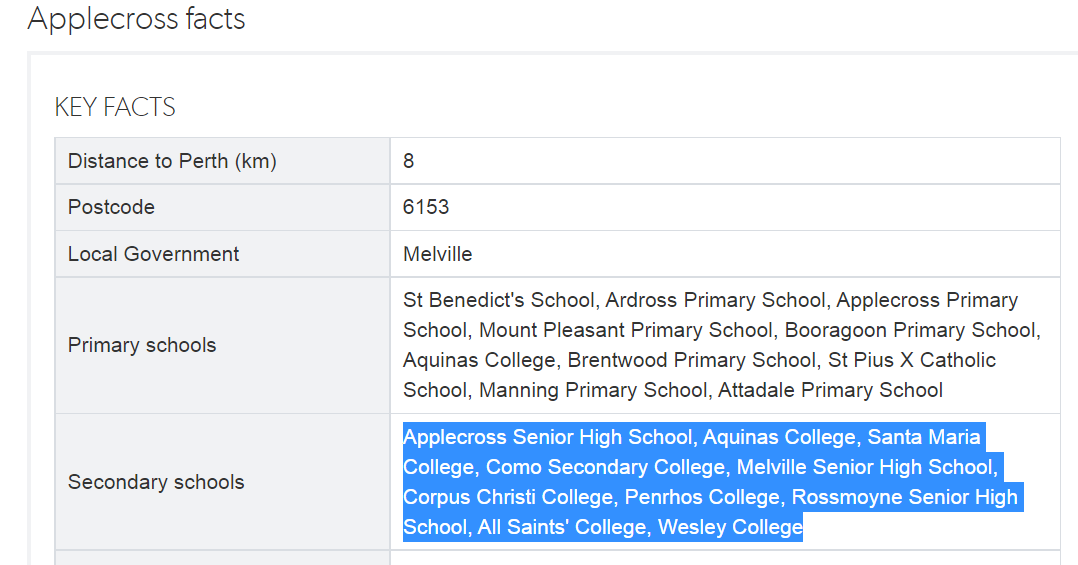
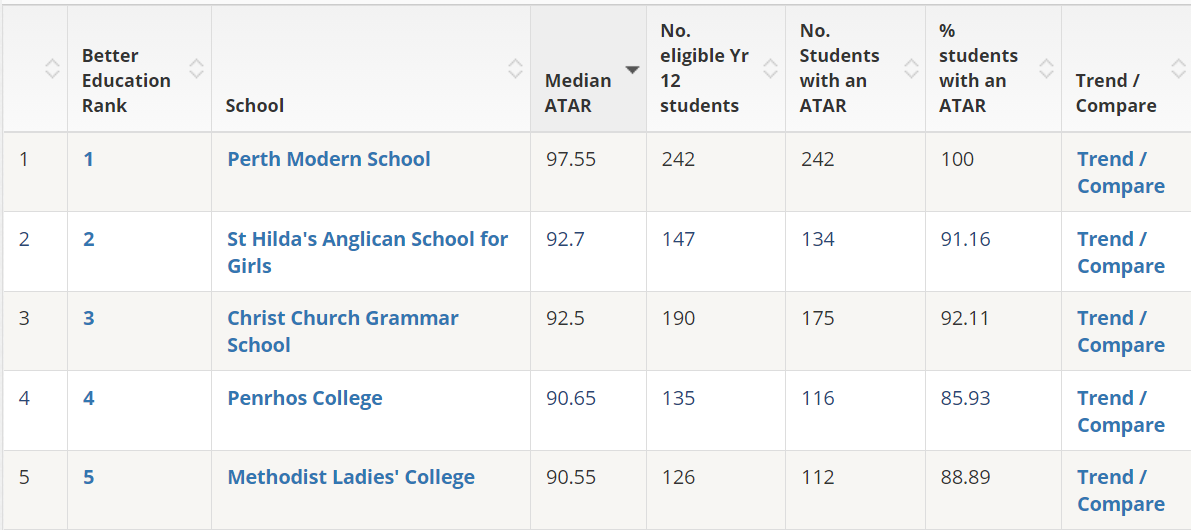
* **Tue 31/8:**
  + Determine project proposal and data sources
  + Commence EXTRACTION
    - Data source 2 - download CSV file, cleanse data and upload into postgresSQL/Python
* **Thu 2/9:**
  + Complete EXTRACTION
    - Data source 1 - Screen scrape secondary schools
    - Data source 3 - Screen scrape school and ranking
  + Commence TRANSFORMATION
    - Dataset 1
    - Dataset 2
* **Mon 6/9:**
  + Complete TRANSFORMATION
    - Dataset 3
    - Dataset 4
  + Complete LOAD
    - SQL Schema
    - Load into SQL

## **EXTRACT:**

1. **Data source 1** – Alphabetical List of Western Australian Schools (XLSX)
   1. **Method** – CSV download - pandas
   2. **Purpose** – obtain a list of WA schools and a flag to see whether it is government or non-government, as well as a flag to determine how many children are in the Yr 12 cohort
   3. <http://det.wa.edu.au/redirect/?oid=SiteProxy-id-19622149&title=Alphabetical+List+of+WA+Schools+xlsx&skip=true&launch=true>

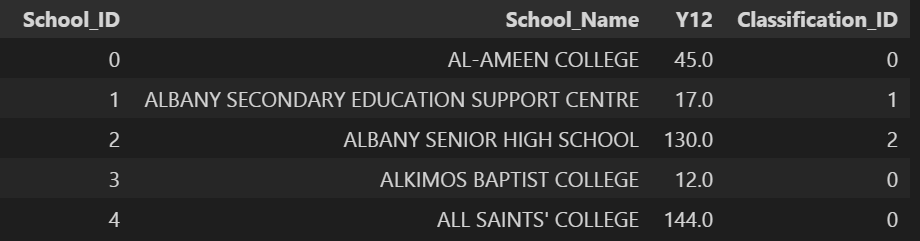


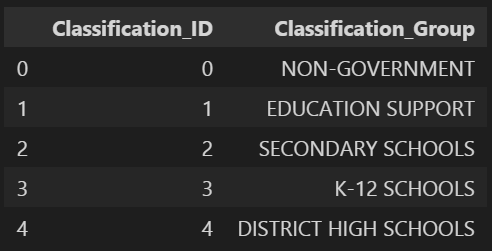
1. **Data source 2 –** REIWA Applecross suburb profile, secondary schools
   1. **Method** - Screenscrape – beautiful soup
   2. **Purpose** - Extract a list of all the secondary schools in the Applecross area
   3. [**reiwa.com - Suburb profile for Applecross**](https://reiwa.com.au/suburb/applecross/)
   4. 
2. **Data source 3** – WA School Ranking – 2020
   1. **Method** – Screenscrape - pandas
   2. **Purpose** – obtain a list of WA School Rankings so the user can merge on the school information to determine the top ranking public schools
   3. <https://bettereducation.com.au/results/wa/wace.aspx>
   4. 

## **TRANSFORM:**

1. **Dataset 1:** wa\_schools\_y12 (new\_WASchools\_df2)
   1. Data source: 1
   2. Transform:
      1. Keep only relevant columns for cleaner dataset
      2. Remove instances where Y12 isNull for relevance
      3. Create School ID Primary Key to make it easier for enduser to merge
      4. Rename columns for consistency across all datasets
      5. Merge on Classification ID foreign key from dataset 2
   3. Columns:
      1. School\_ID - primary key
      2. School\_Name - varchar(100
      3. Y12 – int
      4. Classification\_ID foreign key



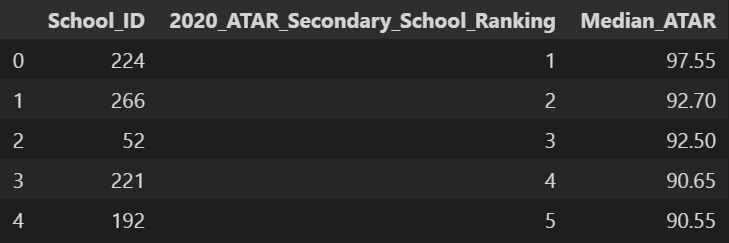
1. **Dataset 2:** wa\_schools\_classification (classification\_df)
   1. Data source: 1
   2. Transform:
      1. Same as above
      2. Create separate dataset with only unique classification group, so it can be updated easily in the future
      3. Create classification ID primary key
   3. Columns:
      1. Classification\_ID – primary key
      2. Classification Group – varchar(100)



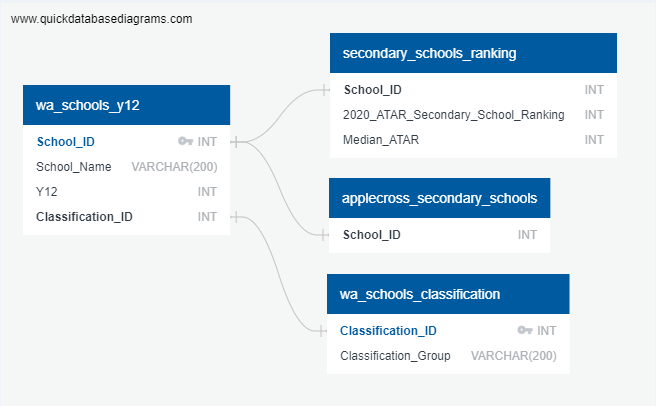
1. **Dataset 3:** applecross\_secondary\_schools (applecross\_df2)
   1. Data source : 2
   2. Transform:
      1. Rename column name for consistency across all data
      2. Make uppercase for ease of merging
      3. Strip school\_name for ease of merging
      4. Merge on classification id foreign key from dataset 2
   3. Columns: Secondary School, Suburb
      1. Note – Suburb is hardcoded as ‘Applecross’



1. **Dataset 4:** secondary\_schools\_ranking (secondary\_schools\_df2)
   1. Data source: 3
   2. Transform:
      1. Rename columns for easier use
      2. Transform school name to upper string for easier merging
      3. Correct any bad data where school name is not consistent with dataset 1
      4. Merge on school id foreign key from dataset 1
   3. Columns: S
      1. School\_ID = int
      2. 2020\_ATAR\_Secondary\_School\_Ranking - Int
      3. Median\_ATAR – Int



## **LOAD:**



1. **SQL Databases**
   1. Schools\_db
      1. Datasets x 4

|  |
| --- |
| Python Outputs |
|  |
|  |
|  |
|  |