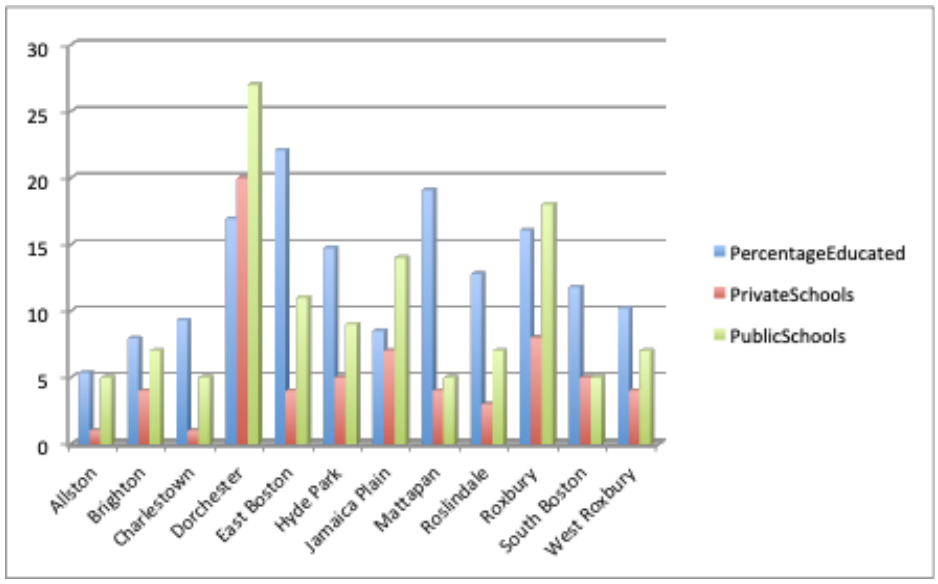
HubLyfe PM5
Hypothesis, Data Sources, and ETL Workflow Descriptions:

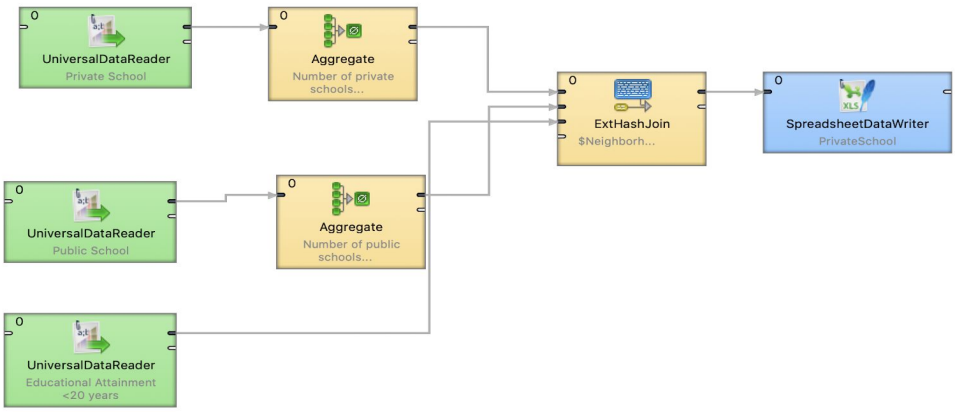
1) Private Schools Hypothesis

Private Schools in Boston Data:
This is a dataset for the private schools in Boston by neighborhood.

Link:https://data.boston.gov/dataset/non-public-schools/resource/2dd752ec-6291-4ee4-aec0-17af0fdc7ff1

Prediction: The educational population (<20 years) will be directly proportional to the total number of private schools and public schools





ETL: Aggregation to obtain number of Private Schools by Neighborhood (i), Aggregation to obtains number of Public Schools by Neighborhood (ii), extracting total population for age range '0-10 years' & '10-19 years' along with total population by educational attainment for 'Less than High School' and 'High School and GED' by Neighborhood (iii) from our existing 'Demographic', 'AgeData' & 'EducationalAttainment' tables. Performing inner join i, ii, iii on Neighborhood name to obtain the output.

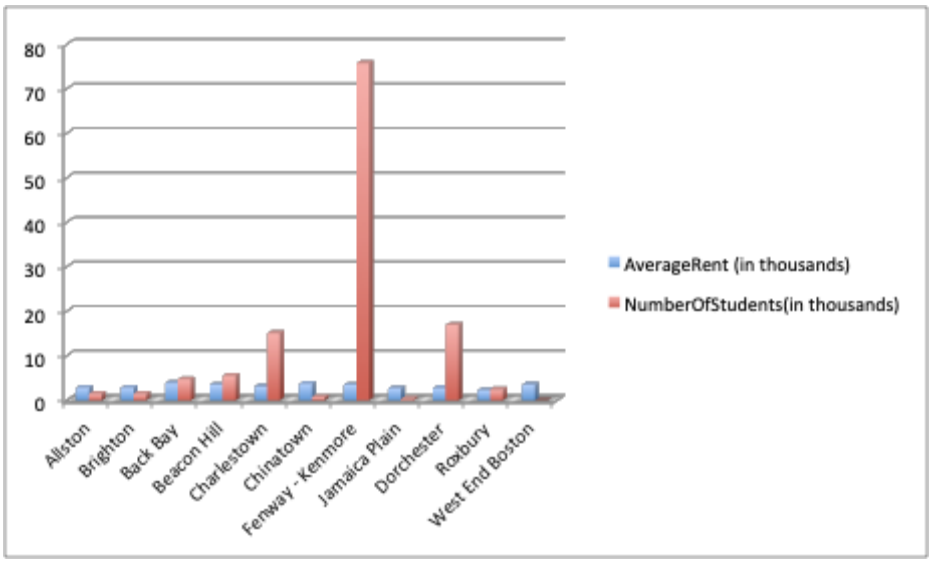
Conclusion: We found that the educated population (public and private school students under 20 years old) in neighborhoods is not strongly correlated to the number of schools in that neighborhood. This is most likely because students travel to different neighborhoods to attend different schools.

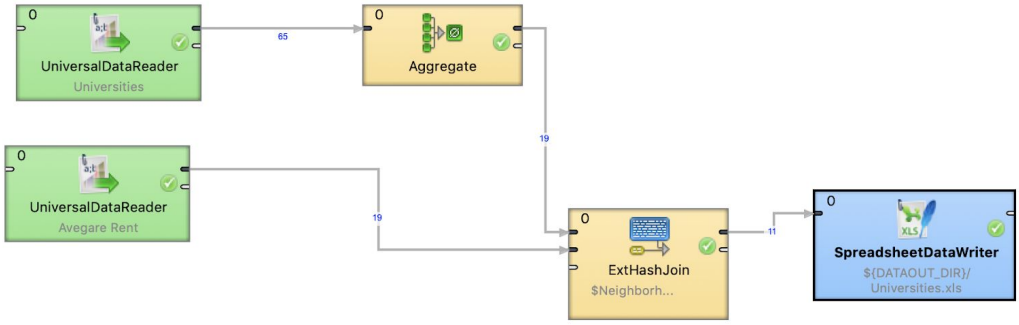
2) **Universities in Boston (hypothesis 1)**

Universities in Boston Data:
This is a database for all colleges and universities within Boston, including neighborhood location

location.https://data.boston.gov/dataset/colleges-and-universities/resource/208dc980-a278-49e3-b95b-e193bb7bb6e4

Prediction: Average rent (irrespective of occupancy type) would be directly proportional to the number of students in the neighborhood



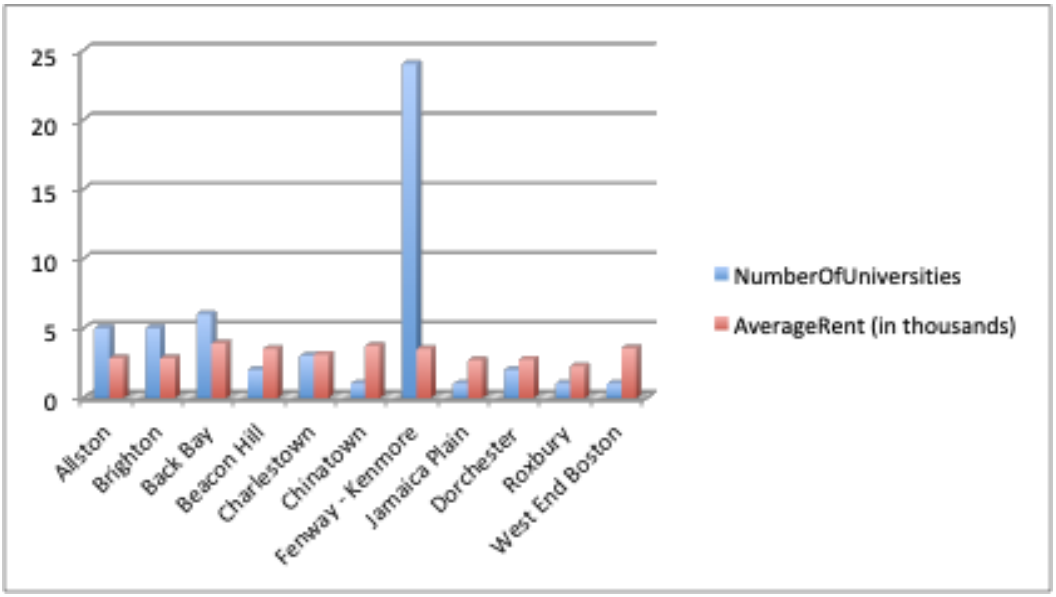


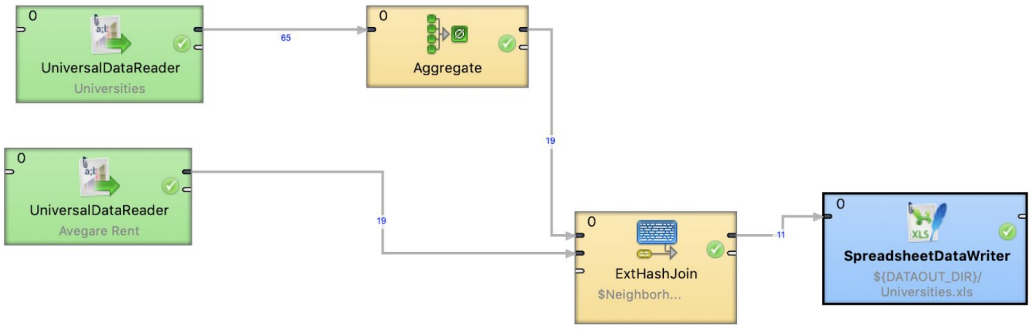
ETL: Aggregating Universities data to obtain number of students by Neighborhood(i), extracting existing average rent by Neighborhood (ii) from our existing ‘Rent’ table. Performing inner join on Neighborhood name to obtain ‘Average Rent’ and ‘Number of Students’ by Neighborhood.

Conclusion: We discovered that there is not a direct positive relationship between the average rent prices in a neighborhood to the number of students in that neighborhood. We can derive that Boston neighborhood rent prices are mostly independent of students living in that area.

3) **Universities in Boston (Hypothesis 2)**

Prediction: The number of universities in a neighborhood will have a positive correlation to the rent price due to increased demand for part time tenants (students).





ETL: Aggregating Universities data to obtain number of universities by Neighborhood(i), extracting existing average rent by Neighborhood (ii) from our existing ‘Rent’ table. Performing inner join on Neighborhood name to obtain ‘Average Rent’ and ‘Number of Universities’ by Neighborhood.

Conclusion: For our second hypothesis using the university data, we found that the number of universities in an area also does not strongly correlate to the average rent price. However, we cannot say that universities do not affect the rent price in a neighborhood. It is only one of the factors that drive the rent price up.

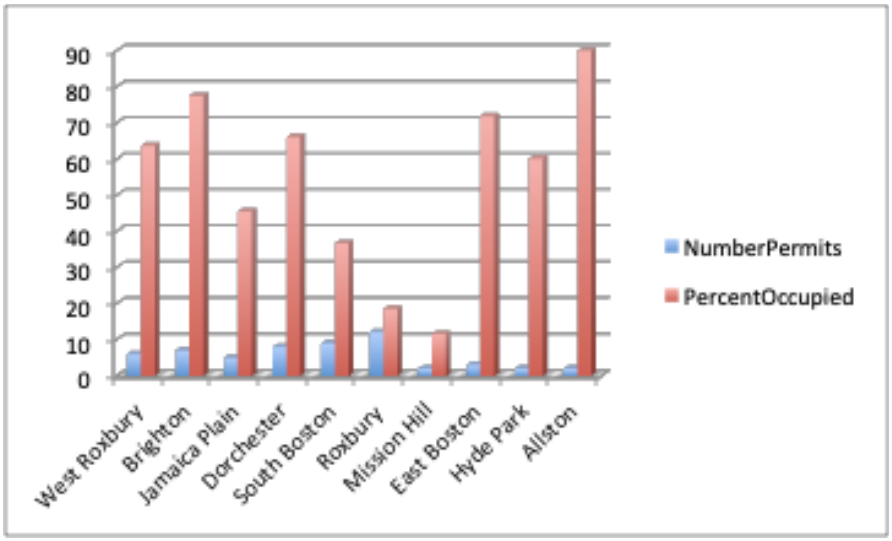
4) **Building Permits Hypothesis**

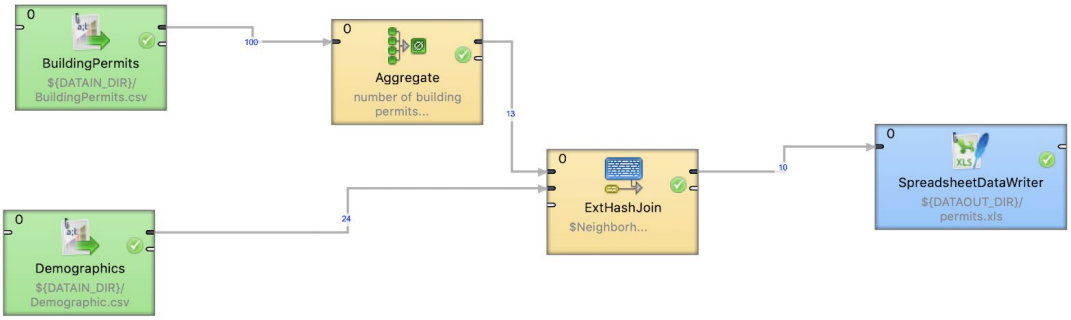
Building Permit Data:

This is a list of all the issued building permits by the city of boston in recent history (past few years) and the various project type, location (by neighborhood) and project owners that are conducting the respective building project.

Link:https://data.boston.gov/dataset/approved-building-permits/resource/6ddcd912-32a0-43df-9908-63574f8c7e77

Prediction: We predict that the number of building permits will correlate to a larger number of renters in a neighborhood due to the amount new construction and renovations by landlords and management companies.





ETL: Aggregated Building Permits data to obtain the number of building permits by Neighborhood (i), extracted demographics data from our existing 'Demographic' table(ii). Performed inner join on the above to obtain 'Number of Permits', 'Number of Renter Occupied Units' and 'Number of Total Units' by Neighborhood.

Conclusion: We found that there is no relation between the number of building permits to the number of renter occupied homes in a given neighborhood.

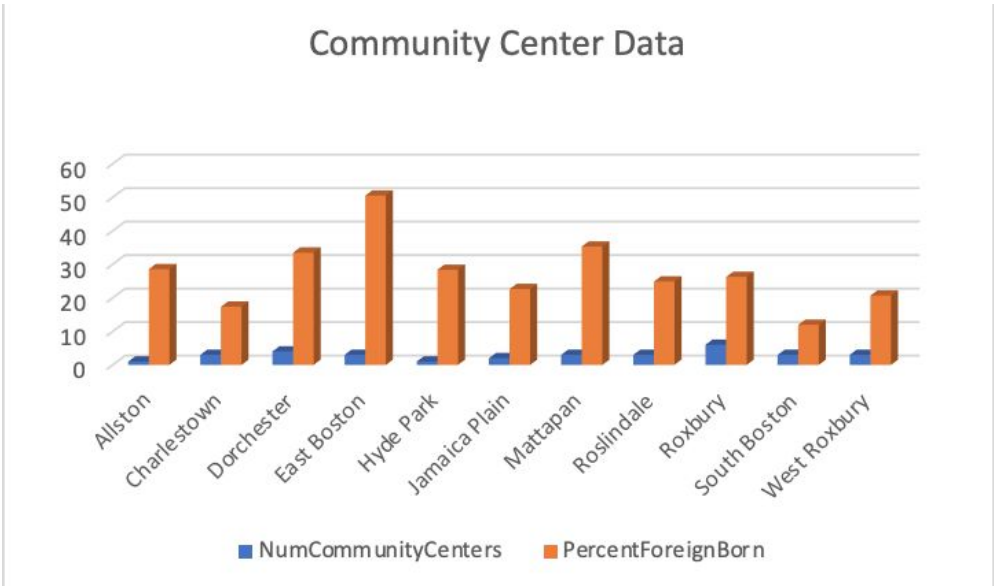
5) **Community Centers Hypothesis**

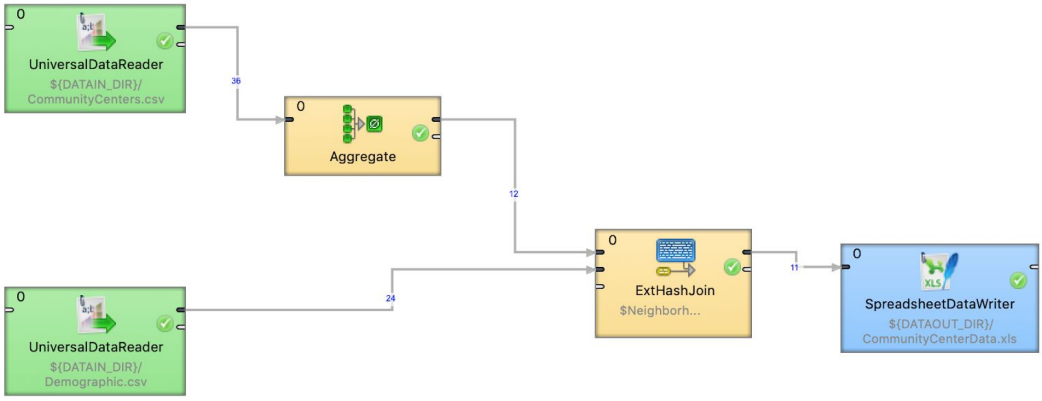
Community Center Data:

This is a list of all community centers and their respective neighborhood locations in boston.

Link:https://data.boston.gov/dataset/community-centers1/resource/f793c76f-145c-4755-9f07-7b73ac2d724b

Prediction: We predict that the number of community centers will be higher in neighborhoods with larger populations, and in particular a larger percentage of immigrant families.





ETL: Aggregating community centers on neighborhood to find the number of community centers per neighborhood, (ii) extracting neighborhood data, population data, and foreign born population data from our ‘Demographic’ table. (iii) Then inner join on neighborhood to find all neighborhoods with community centers and their respective population and foreign born populations.

Conclusion: We found that there is no strong correlation between the number of community centers in a neighborhood to the percentage of foreign born residents. Our hypothesis was incorrect.

ETL workflows:

1. Aggregate: In the context of our workflow, we use aggregate to get group information for eg, by neighborhood or zip code.
2. ExtHashJoin: This is used to inner join data extracted from new data sources and our previously existing data on a common key.