Team Rad Cats
PM5 – App Report

External Data Sources and Hypothesis

Proposed Capital Investments from 2011-2016: contains projected expenditures in Seattle by department, project, budget control level, and year. https://data.seattle.gov/Finance/2011-2016-Proposed-Capital-Investment-Program-CIP-/9689-kxj4

We plan to use the data to see if there is any correlation between the progression of housing prices throughout Seattle from 2013-2016 and any increases or decreases in proposed spending. We especially want to see whether there is a correlation between proposed investments in certain areas over others, like investment in forests vs library major maintenance.

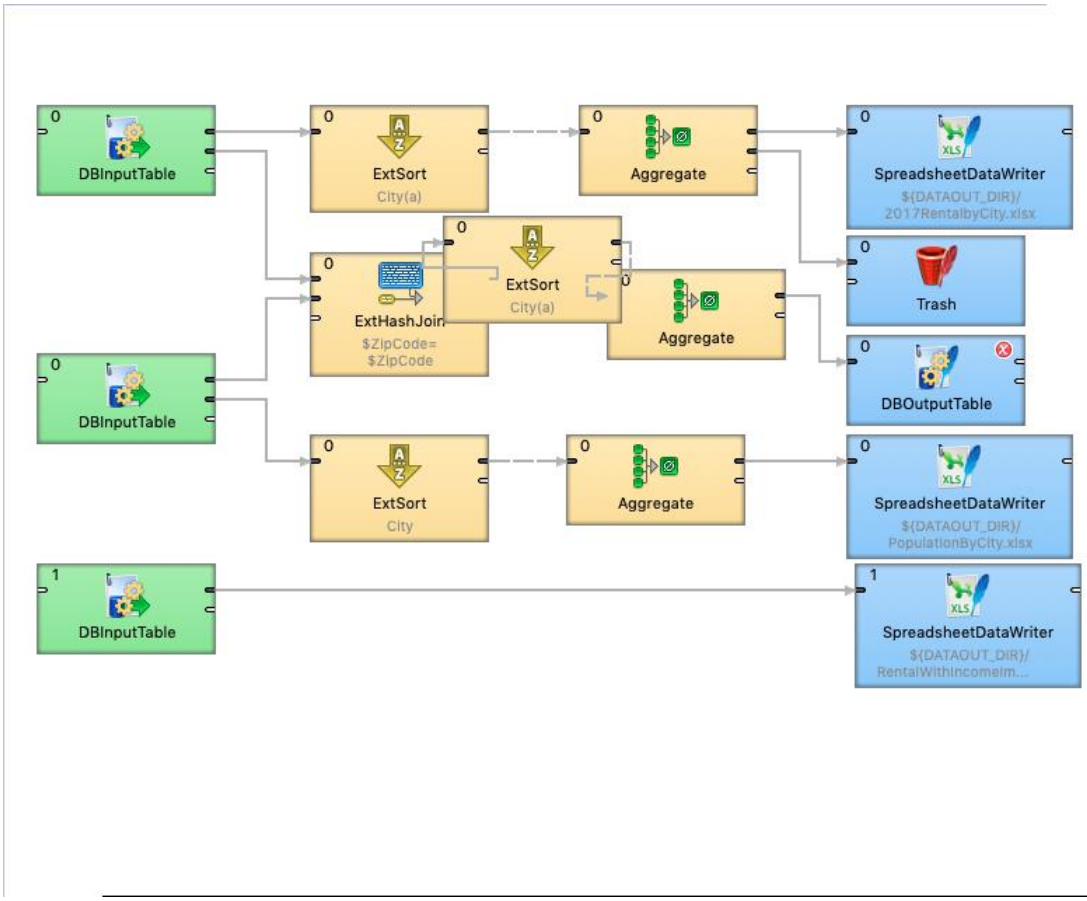
Metropolitan Area Gross Domestic Product in Seattle, Tacoma, and Bellevue: Contains the total gross domestic product from 2001 to 2017 in various industries as noted by the Bureau of Economic Analysis.
https://apps.bea.gov/
We plan to use it to see if there are any correlations between the fluctuations in GDP and the housing and rental prices. And if so, how much. Also want to see if there is a correlation between this and crime prevalence.

Seattle Cultural Space Inventory: All cultural spaces that have existed in Seattle from 1891 on.
https://data.seattle.gov
One hypothesis is that there may be a correlation between median housing/rental prices and the types of cultural spaces that start or are able to continue in each neighborhood. Our hypothesis is that the higher the price is, the more minority centric spaces disappear from a neighborhood.

WA Income by Zipcode: Contains the average income by zipcode for cities in Washington as well as the national rank for the city.
zipatlas.com/us/wa/zip-code-comparison/median-household-income.htm
We wanted to see if there was any correlation between the income of the residents living in the city and what the rental prices were. We hypothesized there would be a correlation; the higher the income, the higher the rental prices in the city.

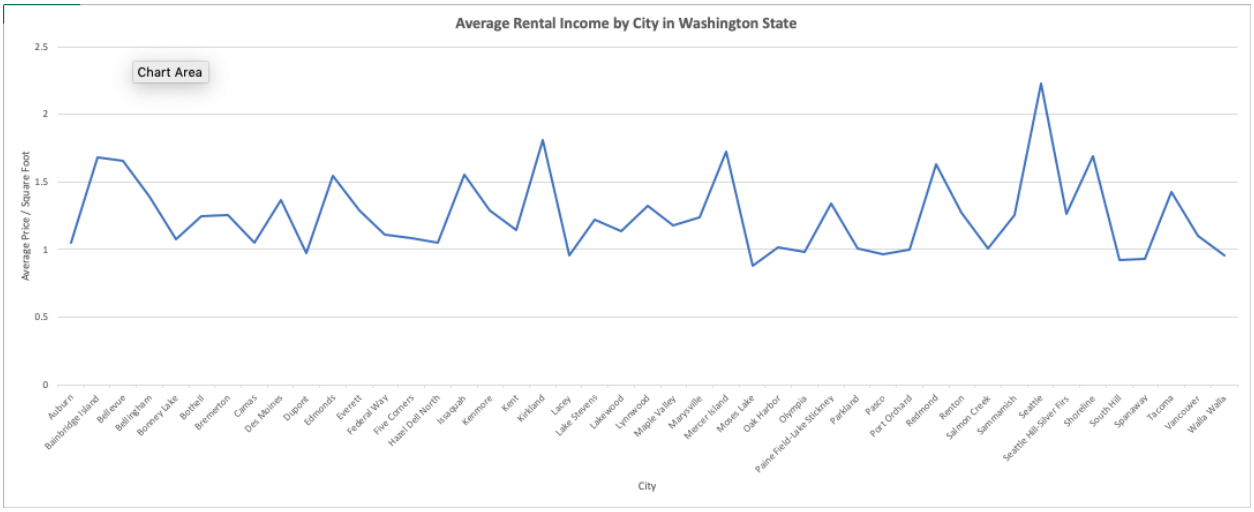
ETL Workflows

To investigate these questions we created two workflows. For questions concerning income and rental prices, we used this workflow:



For our first chart, we took our initial database that combined historical rental data, rental forecast data, and zipcodes. The data from our current database outputs the zipcode, the city, and the price for 2017. This data runs through extsort, which sorts it alphabetically by the city and then outputs this data. Then it runs through an aggregator that gets the average of the price for each city in 2017. It's then output to the spreadsheet data writer that writes it to a new spreadsheet.

The result of this spreadsheet is the following graph:

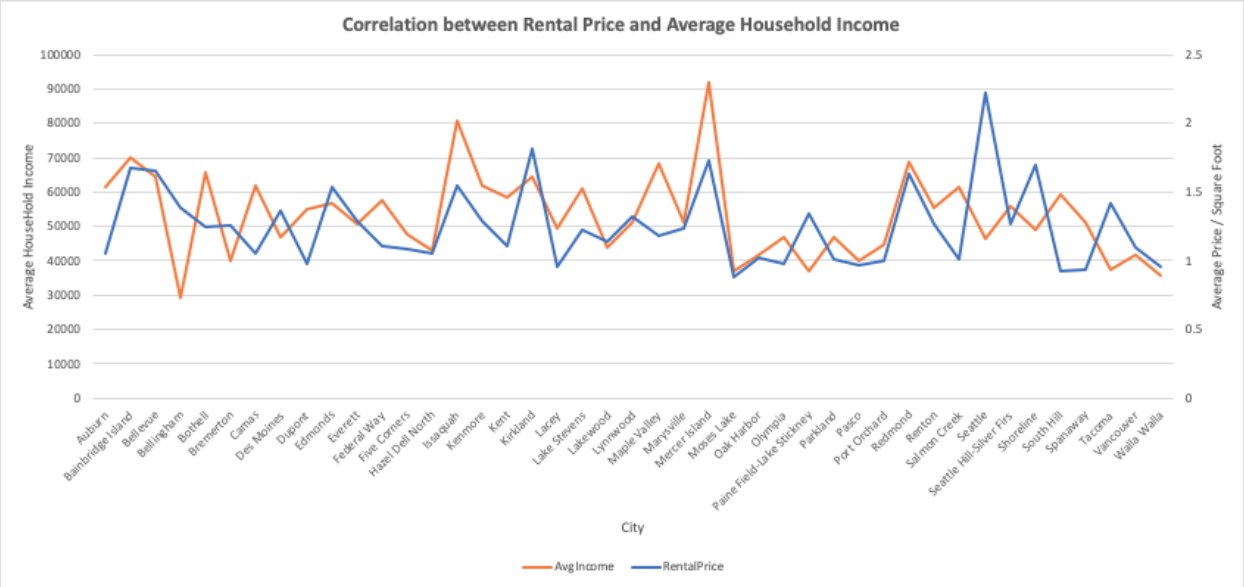


Our initial hypothesis is that Seattle would have the highest rental costs in the region. As shown above, our hypothesis was proven.

Our goal is to create a more nuanced view of the rental and housing market in Seattle by incorporating more information. In the future, we wanted to extend our application to include more cities than Seattle and this chart and the information contained within goes towards that. With the new information, we could do a city to city comparison in addition to our neighborhood to neighborhood comparisons.

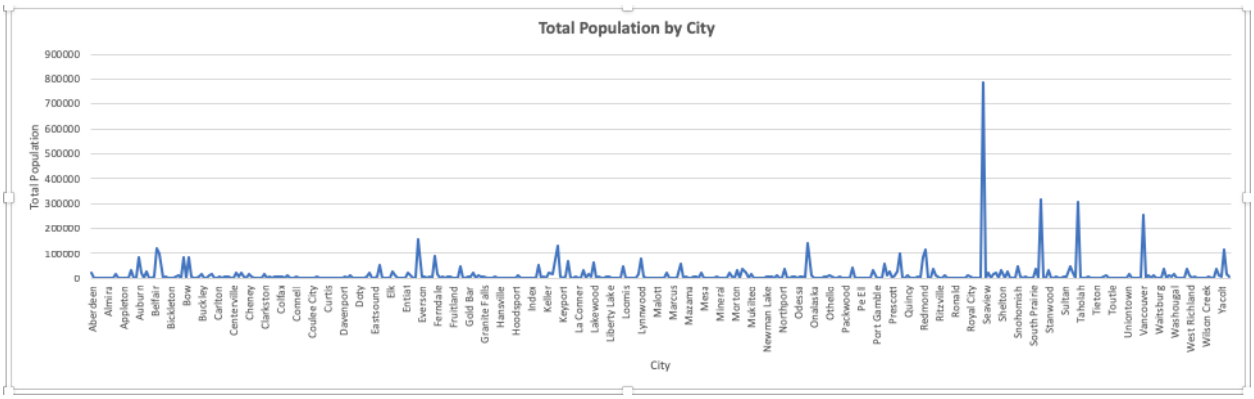
The second graph results from the same database used in the first workflow. It is exthashjoined with a second database. The second database contains a dataset of average income by zipcode in Washington state. In exthashjoin, the databases are joined by the zipcode and outputs the city, rental price, total population, and average income. We then send the output to ExtSort, which sorts the output alphabetically by city. It is then sent to the aggregator, which averages the rental price for each city, sums the total population, and get the average income of each city. This is output to a database table.

This database table is then written to a spreadsheet and results in the following chart:



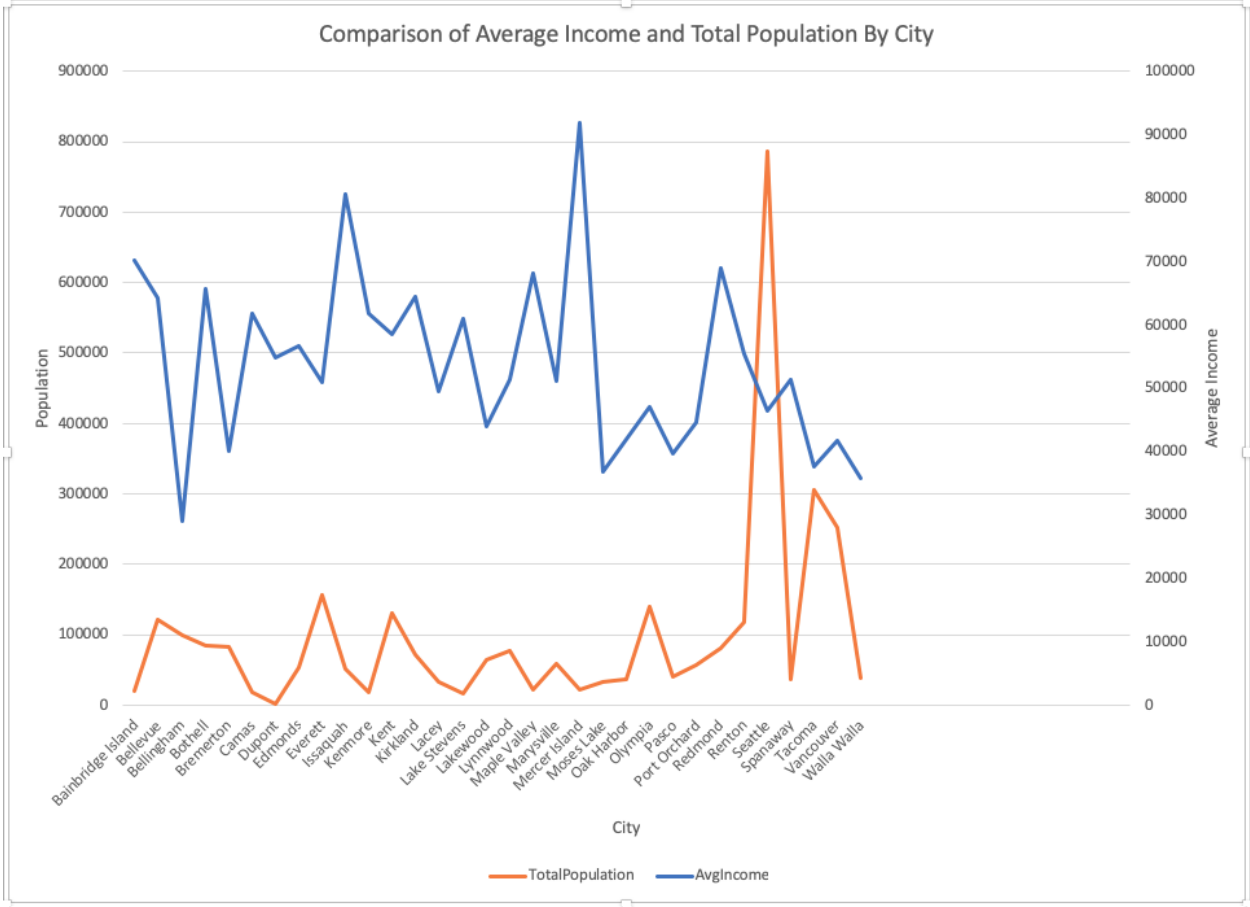
This chart shows the correlation between the rental price and the average household income for cities in Washington state. Our initial hypothesis was that there would be a strong correlation between the two, the higher the income, the higher the rental price. But as shown in our chart, the results were more varied. In some situations, the correlation existed like in Issaquah and Mercer Island. But in other situations, like Bellingham, Paine-Field-Lake Stickney and Seattle the rental prices dropped in comparison to the rising income. This chart provides a more complete view of the rental landscape, by not only comparing Seattle to its neighbors by also showing where the best place to live and work are. It allows the users to see how much of a gap there is between rents and the max they could make in their city and determine whether it is worth living there or if they'd rather move.

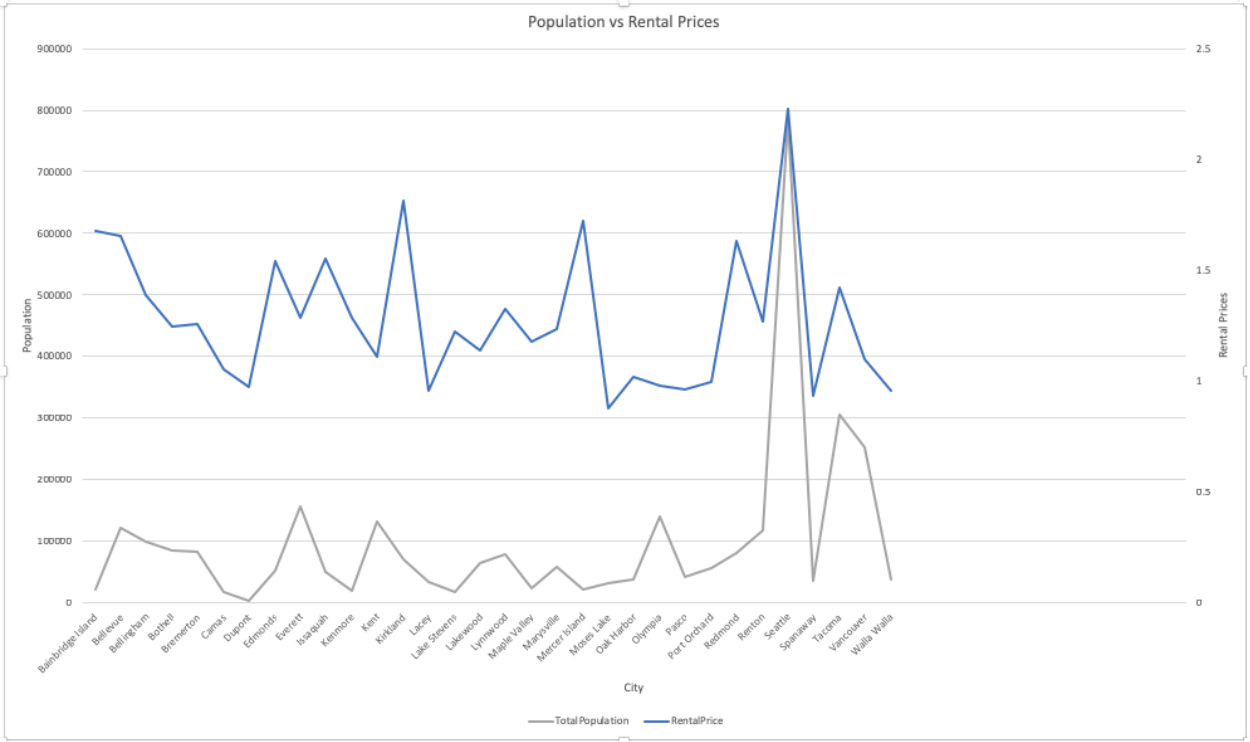
The third chart from this ETL workflow came from the second database we created from the WA average income by zipcode data. We took that database and sent it to ExtSort. We sorted the data by city in ExtSort and then sent it to Aggregation which gave us the average income for each city and the total population for each city. This resulted in just the city, the total population for each city, and the average income for each city. We then sent this to SpreadsheetDataWriter, which created a spreadsheet of the data that leads to the following chart.



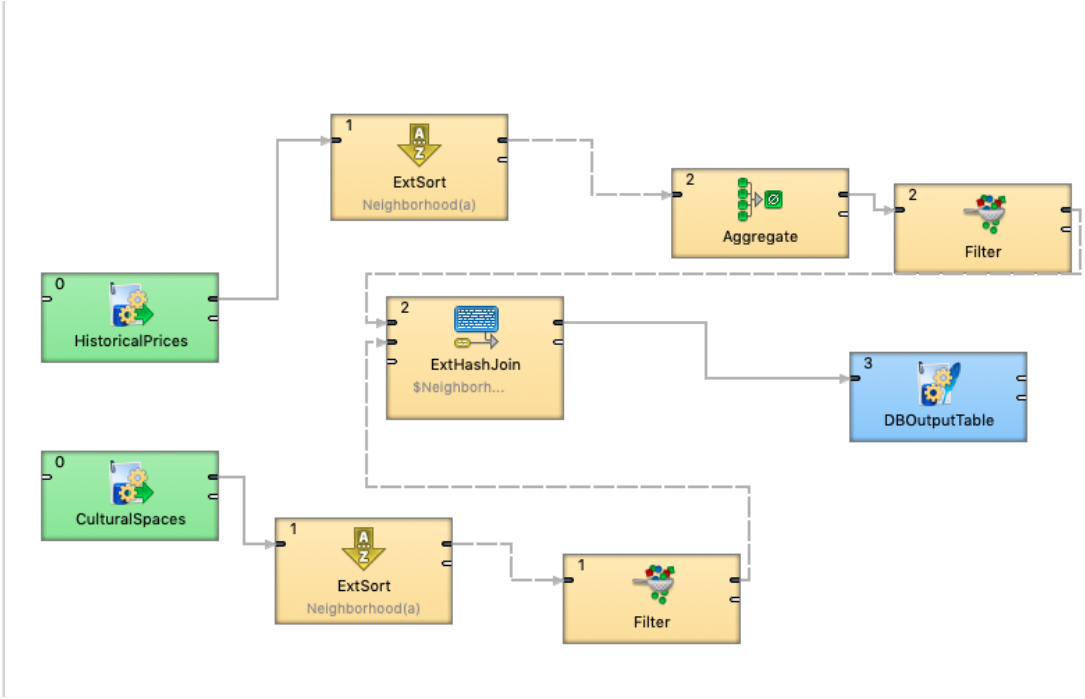
This chart shows the population of each city in Washington state. Our initial hypothesis was that Seattle would be one of the most populated cities and as seen in this chart, our hypothesis was correct. One aspect of choosing a place to live in is how many people will be around you. This chart gives the user a way to visualize that information, and combined with the other charts showing average income, and rental prices per city, it provides the user with a more comprehensive view of what living in a city would be like.

Being able to see all the correlations between our data is something potential customers want because they want a complete view of a neighborhood and we want to give them the ability to do so. To that end, we've created two additional graphs that further show the correlation between population, average income, and rental prices:


Comparison of Average Income and Total Population By City
Population
Average Income
City
TotalPopulation
AvgIncome
Bainbridge Island
Bellevue
Bellingham
Bothell
Bremerton
Camas
Dupont
Edmonds
Everett
Issaquah
Kenmore
Kent
Kirkland
Lacey
Lake Stevens
Lakewood
Lynnwood
Maple Valley
Marysville
Mercer Island
Moses Lake
Oak Harbor
Olympia
Pasco
Port Orchard
Redmond
Renton
Seattle
Spanaway
Tacoma
Vancouver
Walla Walla


Population vs Rental Prices
Population
Rental Prices
City
Total Population
RentalPrice

Our second ELF workflow:



For this workflow, we took the preexisting database and filter it out so we only had the zipcode, the historical rental prices from 2013-2016, the city, the metropolitan area, and the state. The second database is a table that solely contains information on cultural spaces in Seattle, which is the outside database we incorporated for this project. This database had the center's name, the neighborhood it existed in, the phone number for the organization, the specific demographic and community it dealt with, the year it first occupied the space, whether it closed, and when it closed. We aggregated the cultural database, so we could filter it based on the number of cultural centers that started each year. We then sorted both databases in ExtSort by the neighborhood, and filtered them so that only records that contained a neighborhood would be used. The we ExtHashJoined it using the Neighborhood as a key. The result of this is sent to a database table we have in mySQL to compile the information.