Milestone 5: Data Warehousing

(Wen Liao, Kejian Tong, Houming Leng, Jingyi Mao, Zhuocai Li)

1. Relation between average income and average housing price in each ZIP Code

1.1 Data Sources

Internal Data	External Data
Average Rental Price in Each Zip Code	Average Income / Household in Main Neighborhoods (ZIP Code) of Seattle

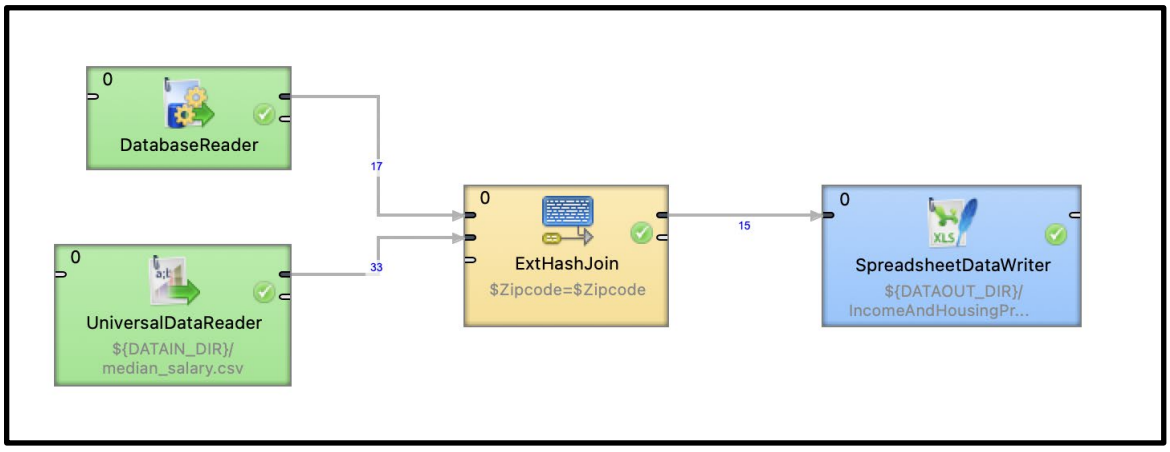
1.2 Hypothesis

As for our expectation of the final results, we assume that the higher the average income in a specific ZIP code area, the higher the rental price it will have.

1.3 Analysis Process

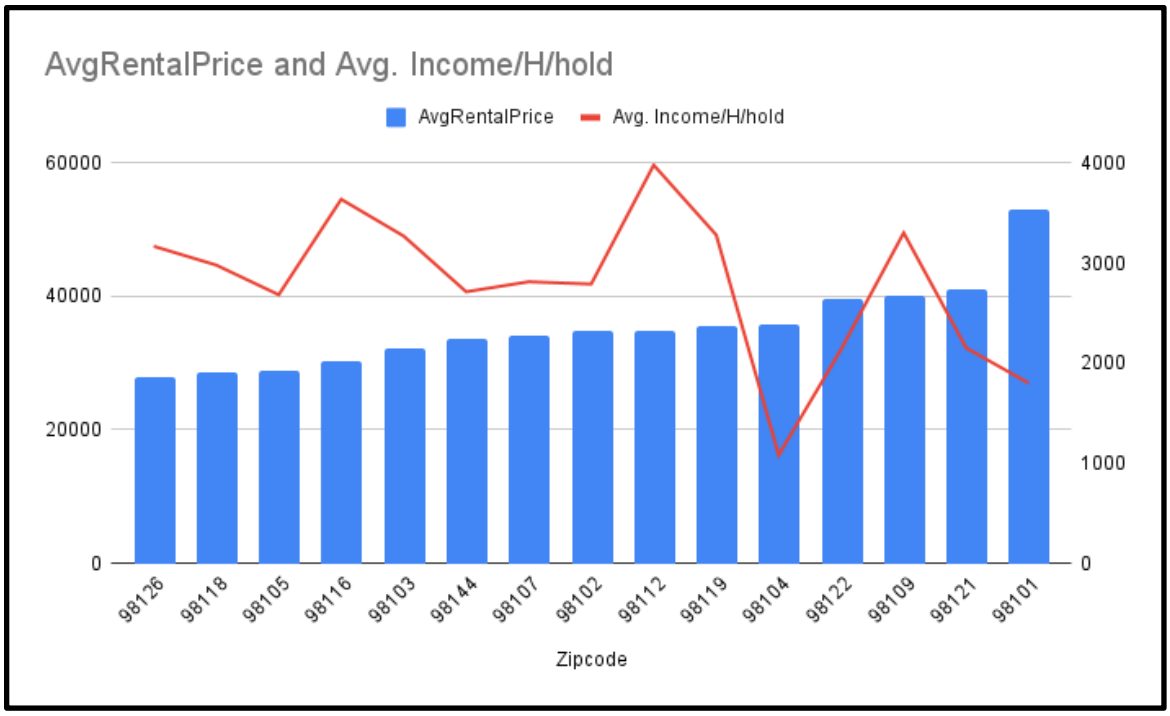
1.3.1 ETL Working Flow

We deployed CloverDX to read multiple data sources, aggregate them with specific data that we need, join them, and push these data to a table that we built previously. At last we output the data into several different spreadsheets in different phases.



pic 1.3.1

1.3.2 Results of analysis on rental price and income



pic 1.3.2

1.3.3 Results of analysis on overall data

As shown in the chart above, the areas with higher average income have lower rental prices. One speculation is that the areas with higher average income have less people needing rental houses.

2. Relation between average rental price and covid vaccine rate in each ZIP Code

2.1 Data Sources

Internal Data	External Data
Average Rental Price in Each Zip Code	COVID Vaccine Rate in the Main Neighborhoods (ZIP Code) of Seattle

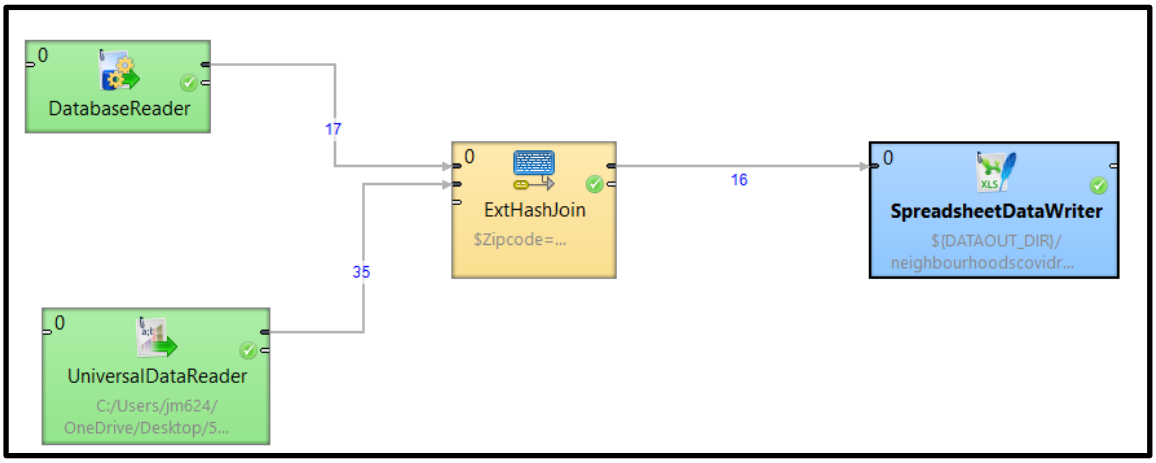
2.2 Hypothesis

With people's higher awareness of the essence of COVID Vaccination, the COVID Vaccine rate of a specific district may be a crucial factor for people to choose to live. the higher average one bedroom rental price in the ZIP code, the higher the vaccine rate.

2.3 Analysis Process

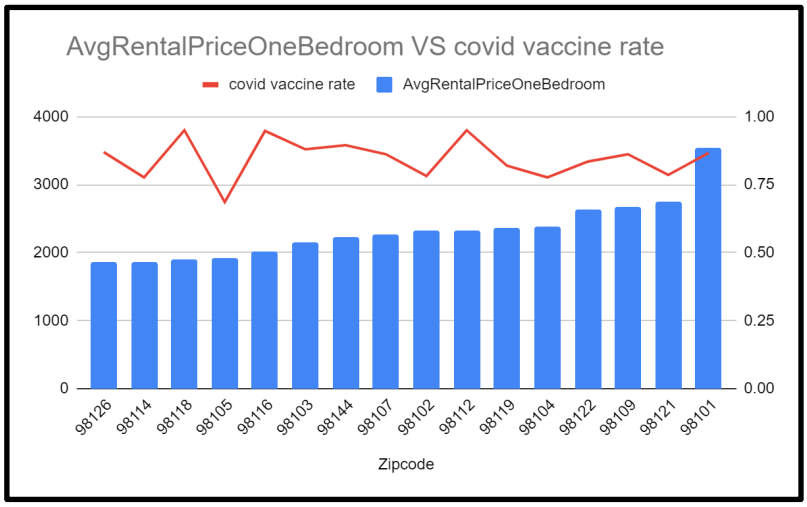
2.3.1 ETL Working Flow

We deployed CloverDX to read multiple data sources, aggregate them with specific data that we need, join them, and push these data to a table that we built previously. At last we output the data into several different spreadsheets in different phases.



pic 2.3.1

2.3.2 Results of analysis on rental price and income



2.3.2

2.3.3 Results of analysis on overall data

The vaccine rate is maintained at a relatively high level (above 0.75) in the greater Seattle area which makes it hard to get a clear understanding of the relation between the price and vaccine rate. The slowest vaccine rate does sit at the relatively low rental price area which is consistent with our hypothesis. We may need more housing data to get a more precise conclusion.

3. Relation among parks amount which might influence pets amount

3.1 Data Sources

Internal Data	External Data
Seattle Parks Open Data	Seattle Licensed Pets Data

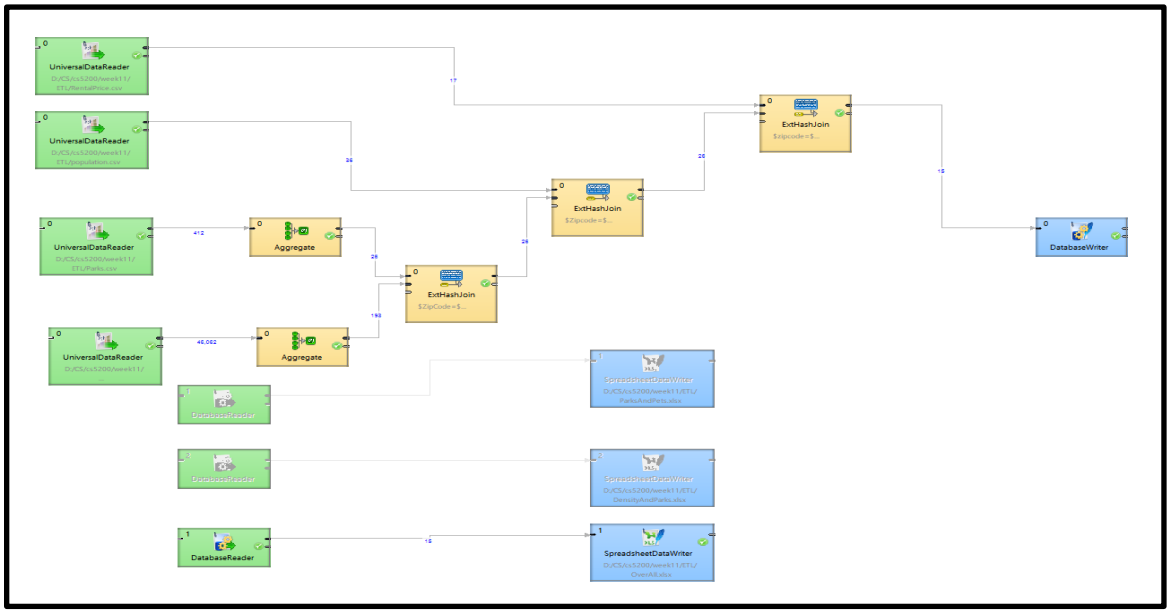
3.2 Hypothesis

The rental price strongly positively affects the rental prices according to our previous studies. This can be caused by the population density of a specific area, which is also influenced by the people's inner intention to live in an area with better living conditions, for example with more parks for them to perform outer-space activities like petting. The number of pets can vividly show that.

3.3 Analysis Process

3.3.1 ETL Working Flow

We deployed CloverDX to read multiple data sources, aggregate them with specific data that we need, join them, and push these data to a table that we built previously. At last we output the data into several different spreadsheets in different phases.



3.3.1

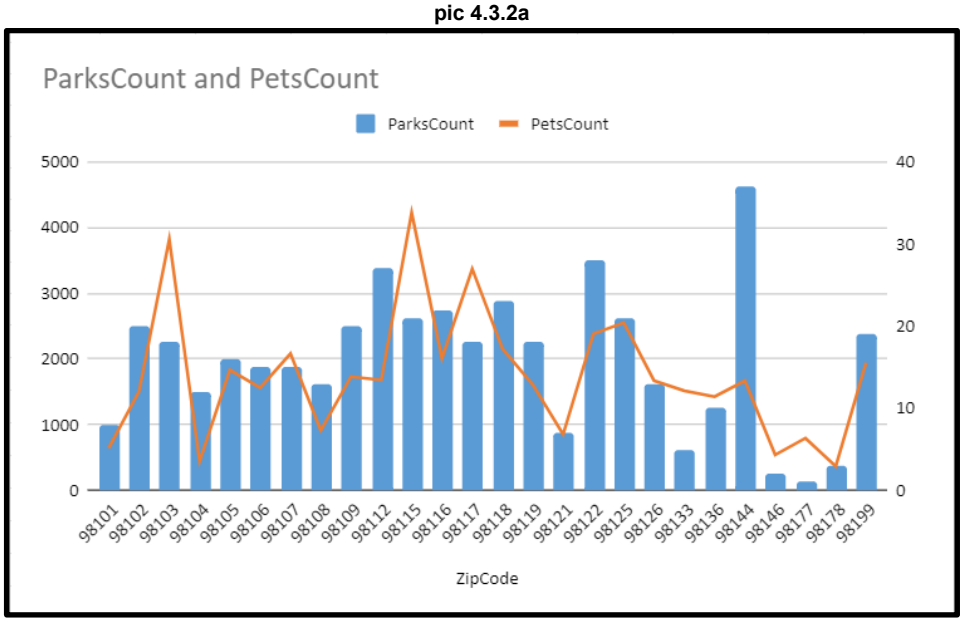
We chose three different phases to analyze the final relation and performed them in chronological order. We will discuss them step by step as follows.

3.3.2 Results of analysis on relation between #parks and #pets

We gathered Seattle parks data (412 rows in total) and Seattle licensed pets data (46,062 rows in total), taking the count of each component as our study objects and joining with the same ZIP codes. We got 26 rows of joining data at the end.

ZipCode	ParksCount	PetsCount
98101	8	646
98102	20	1504
98103	18	3822
98104	12	440
98105	16	1833
98106	15	1562
98107	15	2082
98108	13	913
98109	20	1731
98112	27	1680
98115	21	4226
98116	22	2008
98117	18	3371
98118	23	2153
98119	18	1609
98121	7	852
98122	28	2382
98125	21	2552
98126	13	1670
98133	5	1516
98136	10	1428
98144	37	1668
98146	2	541
98177	1	795
98178	3	367
98199	19	1944

pic 4.3.2a



pic 4.3.2b

From the table and the chart, we can discover that the number of pets in a specific ZIP code area almost strictly follows the number of parks in this area. With a larger number of parks, there will be a larger number of pets. This might be able to explain with an assumption that parks are providing suitable space for pets like dogs to perform outer-space activities.

4. Relation among external living conditions which might influence average rental price

4.1 Data Sources

Internal Data	External Data
Seattle Parks Open Data	Seattle Population Density Based On ZIP Code
Seattle Rental Prices	

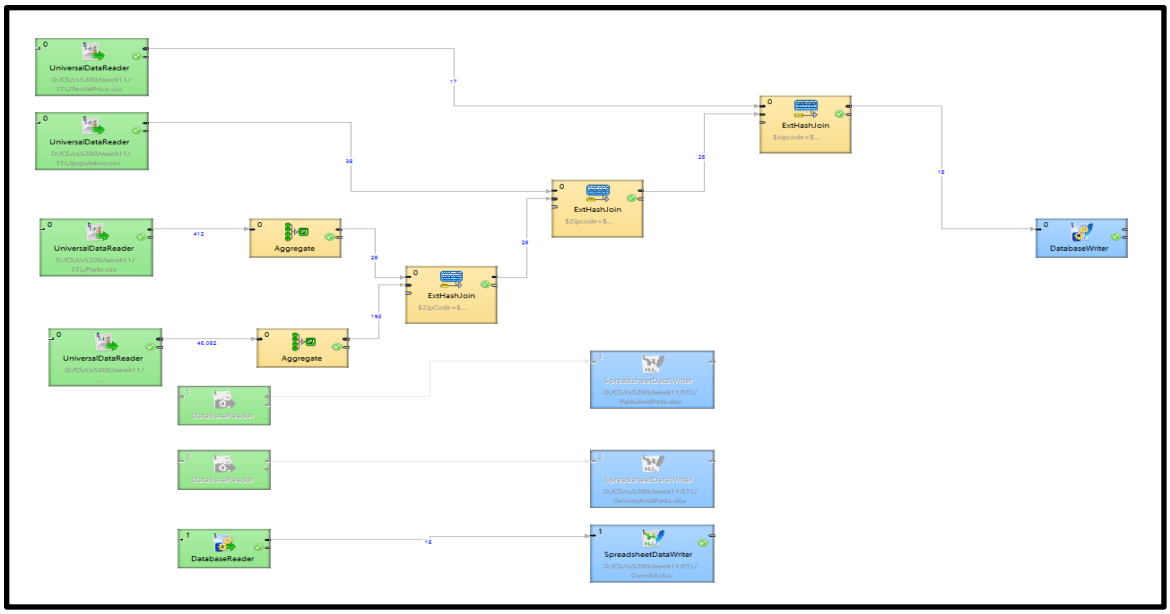
4.2 Hypothesis

The rental price strongly positively affects the rental prices according to our previous studies. This can be caused by the population density of a specific area, which is also influenced by the people's inner intention to live in an area with better living conditions, for example with more parks for them to perform outer-space activities like petting. The number of pets can vividly show that.

4.3 Analysis Process

4.3.1 ETL Working Flow

We deployed CloverDX to read multiple data sources, aggregate them with specific data that we need, join them, and push these data to a table that we built previously. At last we output the data into several different spreadsheets in different phases.



3.3.1

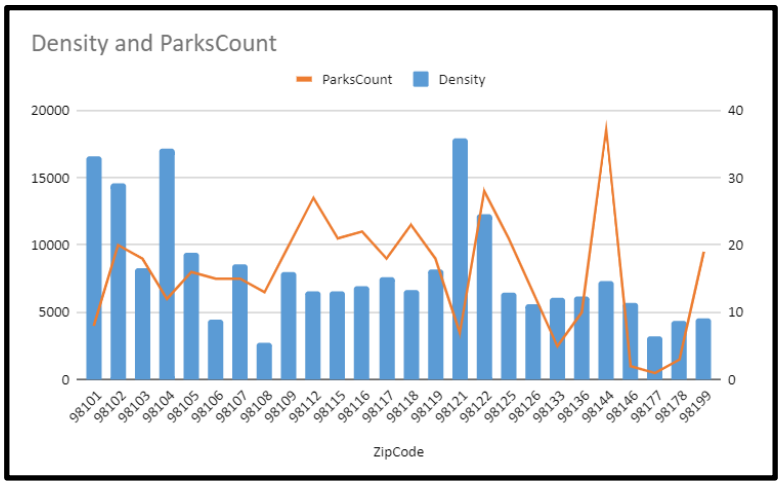
We chose three different phases to analyze the final relation and performed them in chronological order. We will discuss them step by step as follows.

4.3.3 Results of analysis on relation between #parks and population density

We gathered Seattle parks data (412 rows in total) and Seattle population density (26 rows in total), taking the count of each component as our study objects and joining with the same ZIP codes. We got 26 rows of joining data at the end.

ZipCode	Density	ParksCount
98101	16603	8
98102	14594	20
98103	8324	18
98104	17157	12
98105	9394	16
98106	4434	15
98107	8572	15
98108	2764	13
98109	8016	20
98112	6578	27
98115	6603	21
98116	6950	22
98117	7568	18
98118	6697	23
98119	8230	18
98121	17895	7
98122	12332	28
98125	6477	21
98126	5565	13
98133	6042	5
98136	6132	10
98144	7365	37
98146	5745	2
98177	3189	1
98178	4386	3
98199	4551	19

pic 4.3.3a



pic 4.3.3b

As we studied from the table, we found that the number of parks does not strongly affect the population of a specific area. This could be explained by the assumption that parks may not be a priority factor for people to choose an area for them to settle. However, this may also be helpful for us, since our final hypothesis is that parks will affect the rental prices as it could be an essential factor for citizens who require a higher level of living conditions. As a matter of fact, the second study leads to our final study: to gather all data together.

5. Deeper Analysis to Relation between external factors with rental price

5.1 Hypothesis with data source

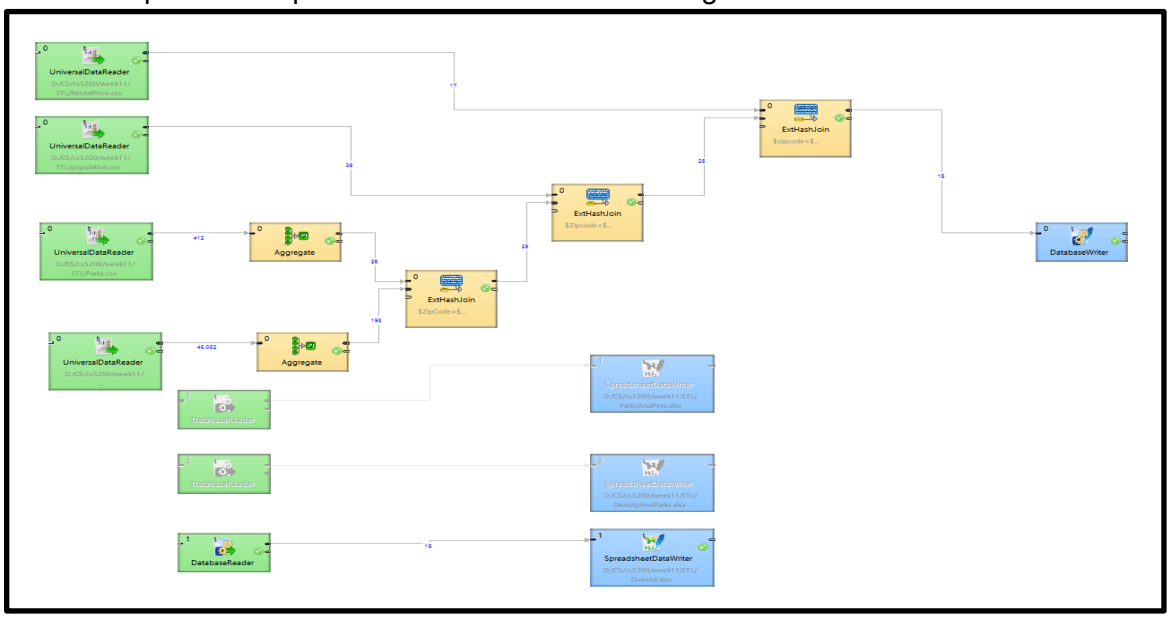
Internal Data	External Data
Seattle Parks Open Data	Seattle Population Density Based On ZIP Code
Seattle Rental Prices	Seattle Licensed Pets Data

The rental price may not be easily concluded to be affected definitely with data mentioned above, since the difference of average rental prices among varied areas is less significant than the difference of density or number of parks. Therefore, we decided to explore other possibilities among these four sources of data.

5.2 Analysis Process

5.2.1 ETL Working Flow

We joined the previous data together and made a new phase of CloverDX working flow in order to explore other possible relations or hidden insight in our data.

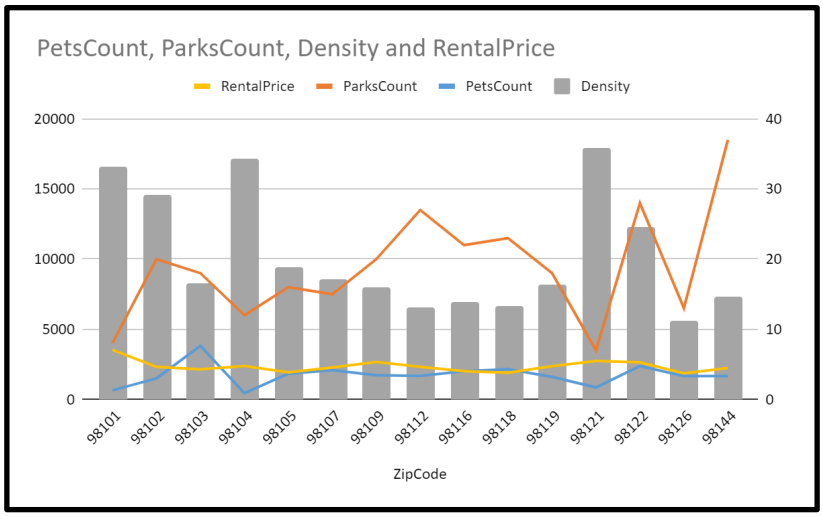


5.2.2 Results of analysis on overall data

We utilized 2 external data and 2 internal data in the final analysis, and joined them with the same ZIP Code. Since not all Zip Code were overlapping, we eventually got 15 rows of integrated data at the end.

ZipCode	PetsCount	ParksCount	Density	RentalPrice
98101	646	8	16603	3535
98102	1504	20	14594	2320
98103	3822	18	8324	2144
98104	440	12	17157	2383
98105	1833	16	9394	1930
98107	2082	15	8572	2275
98109	1731	20	8016	2671
98112	1680	27	6578	2320
98116	2008	22	6950	2017
98118	2153	23	6697	1909
98119	1609	18	8230	2375
98121	852	7	17895	2742
98122	2382	28	12332	2642
98126	1670	13	5565	1859
98144	1668	37	7365	2236

pic 4.3.2a



pic 4.3.2b

As we can see from the table and the chart, the overall data is not following a very strict linear relation or positive relation. However, in some ZIP code areas, with the least population density, there are still a large number of pets and also a relatively large number of parks.

This could be concluded that in in areas of higher population density (with downtown area ZIP codes), the percentage of people who have pets could be lower; however, in areas of lower population density, there are larger percentage of people who choose to pet some animals, and the number of parks could affect their choice of renting housings. Thus, the rental price would be affected due to these considerations.

External Data Source

Seattle Average Income per Household
http://zipatlas.com/us/wa/seattle/zip-code-comparison/median-household-income.html

Seattle COVID Vaccine Rate
https://kingcounty.gov/depts/health/covid-19/data/vaccination.aspx

Seattle Population Density
http://zipatlas.com/us/wa/seattle/zip-code-comparison/population-density.html

Seattle Licensed Pets
https://data.seattle.gov/Community/Seattle-Pets/ez6b-yzed