SI 618 Project Part 1 Report

Using PySpark and Sparksql to Explore the Income and Rental Price Datasets

Motivation

The quality and happiness of living in a certain geographical place are related to the proportion of rent over monthly income. The less the ratio is, the more people can spend on their interest. In this project, I will discover the relationship between household income and the rental price in different counties over the United States. Does a higher household income in a county lead to a higher rental price in that county? I will explore the US income dataset and rental price dataset using Pyspark to find out the average rental price per county and average household income. Besides, I will rank the top10 county with the highest rent-income ratio.

Research Question:

- Which county is the most expensive to live in according to the average rental price?
- Does the top five highest rental price counties have corresponding highest incomes?
- Which county has the worst cost performance in the aspect of rent?

Data Sources

I used two separate datasets described below.

1. Zillow Rent Index Data

Source: https://www.kaggle.com/zillow/rent-index?select=price.csv

The data contains median estimated monthly rental prices in a certain area from 2010 to 2017. It is available on Kaggle and can be downloaded in CSV format. It includes 13,131 records and 81 variables. I decided to use all 13,131 records because I need enough keys to merge the two datasets and there is no missing data in the variables of interest. Each record shows the *city, county, metro, state, population rank and monthly rental price from November 2010 to January 2017*. For this project, I will mainly use county, and January 2017, the most recent monthly rental price of each record.

2. US Household Income Statistics

Source: https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations?select=kaggle_income.csv

The US Household Income Data contains 32,526 records of incomes by household and geographical location. The data is available on Kaggle and can be downloaded in CSV format. It has 19 columns including state_Name, County, City, Place, Longitude, Latitude, Zip Code, mean, median, standard

deviation of household income on a neighborhood scale. For the completeness of the dataset, I will use all 32,526 records and focus on county and median income variables.

Data Manipulation

Step1 Preprocessing and Conversion

After the first examination, the two datasets do not have any missing value or abnormal in the columns that I need, the rent ranges from 518 to 17985 but income ranges from 0 to 300,000 and I choose to exclude those neighborhoods with 0 median income. Next, I convert CSV to JSON format for spark calculation. I open the two datasets in colab, read them using *csv.DictReader* and *json.dump* each line into new JSON files. The source code can be found in the si618_project_czj_part1.py.

Step 2 Catching information and Creating RDDs in Spark

Then, I upload JSON files to cavium, write two functions to catch the information of County, rental price, median income and make the County variables in two datasets match with each other by stripping ' County' in the columns in income dataset to prepare for merging. Apply *map json.loads* to read my datasets into RDDs and use flatMap to apply my function on RDDs.

Step 3 Join two datasets

I decide to use spark SQL to do the join so I need to save my RDD results to tables by using registerTempTable. To make my table visually clearer, I add the column name to each column. I import *StructField, StringType* and *StructType*, create schema string, then parse the schema string using StructField and StringType, use StructType to create the schema and combine my RDD with schema use *createDataFrame.*

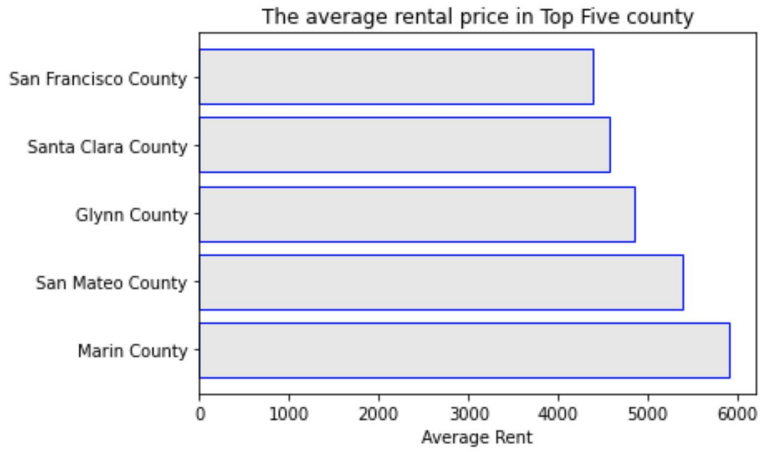
Analysis and Visualization

Task 1 Which County is the most expensive to live in according to the average rental price?

The first task is to find the county with the highest rental price in January 2017. I only keep (county, price) for my analysis. Since there is no abnormal value and missing value, I used all the data points and create a variable count using *mapValue* to count each record in that county
Therefore, my RDD looks like (county, (price, 1)). County is my key and I use reduceByKey to sum up all the rental price in the county. To get the average, I divide the sum of rental prices by the number of records in that county and use sortBy to sort the average from highest to lowest.

The result yields that the county with the highest average rental price 5907 is Marin, CA which locates above San Francisco and has a 247,289 population. The top five counties that has the highest average rental price are Marin CA, San Mateo CA, Glynn GA, Santa Clara CA, and San Francisco CA. Among those, four of them are located in northern California which makes sense because there are a lot of technology companies

that offer high salaries. However, I doubt that Glynn county in Georgia gets into the top three. After further examination, I found Glynn County only have three records in which one of those yield an 11,490 rental price. This explains the reason why the average of that county is boosted extremely.



Task 2 Does the top five highest rental price counties have corresponding highest income?

In the second task, I want to find the average income per household in each county and to check whether the top five highest rental price counties also have higher incomes. I used the median income as my variable than mean because the median can better represent the income level.

I generated (county, (median income, count)) tuples by using mapValues and then sum up income and count by county using reduceByKey, keep the county with at least 2 records using filter, map to get the average, divide the yearly income to monthly by using mapValues and finally, sortBy to sort the average income descendingly.

For Marin County, the average income is 9902, San Mateo has 6246 average income, Santa Clara has 8438 average income and the other two are not included in the income dataset. Among all the 1121 counties in the dataset, the total average income is 6733. Both Marin County and Santa Clara County exceed the total average but San Mateo does not. This result might not be very accurate because the sample size for each county is small, some counties only have one record.

County	Rental Price	Income
Marin	5907	9902
San Mateo	5389	6346
Santa Clara	4562	8438

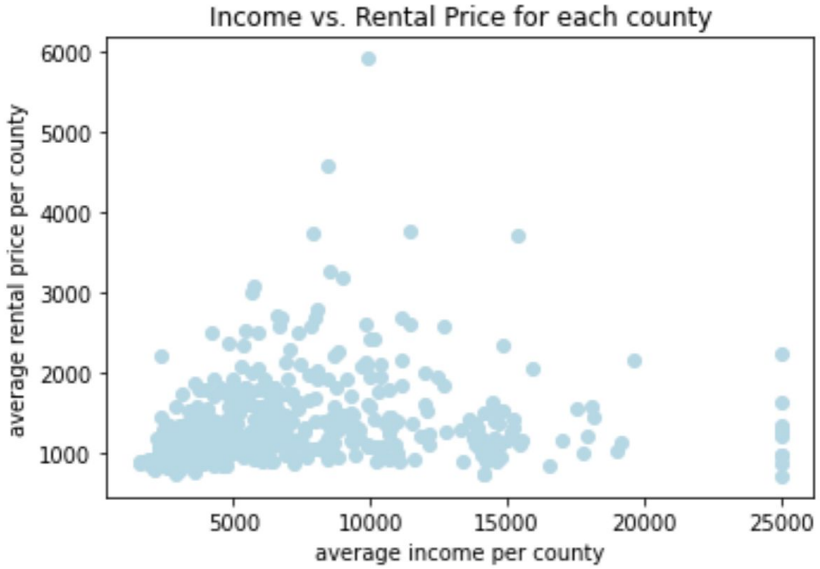
Task 3 Which county has the worst cost performance in the aspect of rent?

Last but not the least, in this task, I will calculate the rent over income proportion of each county. First, I use spark SQL to merge two tables on the county name and write the query to calculate the ratio of the county and order by ratio descending.

```
q1 = sqlContext.sql("select r.county, r.avg_rent, i.avg_income, (r.avg_rent/i.avg_income) as ratio from rent r join income i on r.county = i.county order by ratio desc ")
```

county	avg_rent	avg_income	ratio
Manatee	2199.0	2366.8333333333335	0.9290895007393845
Chesterfield	1438.0	2385.9166666666665	0.6027033634871294
Marin	5906.571428571428	9902.208333333334	0.5964903211224528
Sarasota	2481.8571428571427	4210.458333333333	0.5894505885995331
Bladen	869.4285714285714	1548.7916666666667	0.5613592777779913
Hale	894.0	1605.2083333333333	0.5569370538611291
Schuyler	1371.6666666666667	2479.2916666666665	0.5532494159958322
Lee	1722.1666666666667	3138.805555555555	0.5486694337017798
Santa Clara	4561.818181818182	8438.555555555557	0.5405923030055648
Pickens	1189.25	2224.6666666666665	0.5345744680851064
Lamoille	1563.0	2926.6666666666665	0.5340546697038725
San Diego	2992.714285714286	5636.416666666667	0.5309604421924602
Los Angeles	3063.5454545454545	5771.203125	0.530833066899799
Hopkins	1283.0	2435.472222222222	0.5267972216202653
Broward	1854.0	3576.6	0.5183694011071968
Sumter	898.7	1756.361111111111	0.5116829303009696
Yellowstone	1238.5	2451.125	0.5052781885868733
Kit Carson	1348.0	2698.9583333333335	0.49945194905441914
Davidson	2359.181818181818	4807.416666666667	0.4907379538245041
Wyoming	1208.1	2484.888888888889	0.48617868002146297

From the above table, the top 20 highest ratio counties include Marin, Santa Clara, Los Angeles, San Diego. Many of them are in California and the rest are in Florida and other states. From the scatter plot below, I find that the income and rental price are positively correlated.



Challenge and Limitation

When first attempting to merge the data, only a few rows were successfully merged. Then I found that the value format of the key county is different and it took me a while to match the value from both datasets. There is also limitation such as not enough income data for each county and some counties do not have income data which makes the results inaccurate.