



Data Engineering

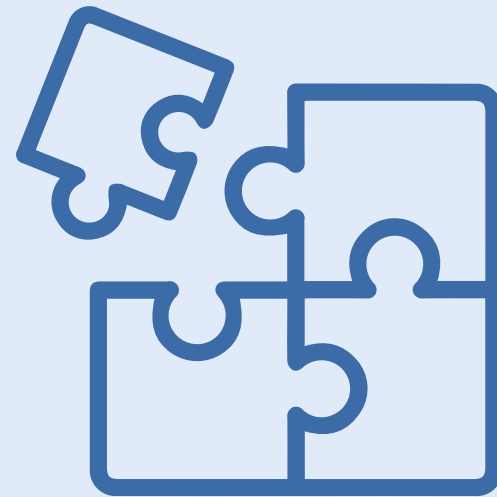
Term Project 2

By: Mayer Mathilde, Pavlova Anna, Karabulut Cansu, Miller Seneca

Project Overview by Stages



Data Extraction



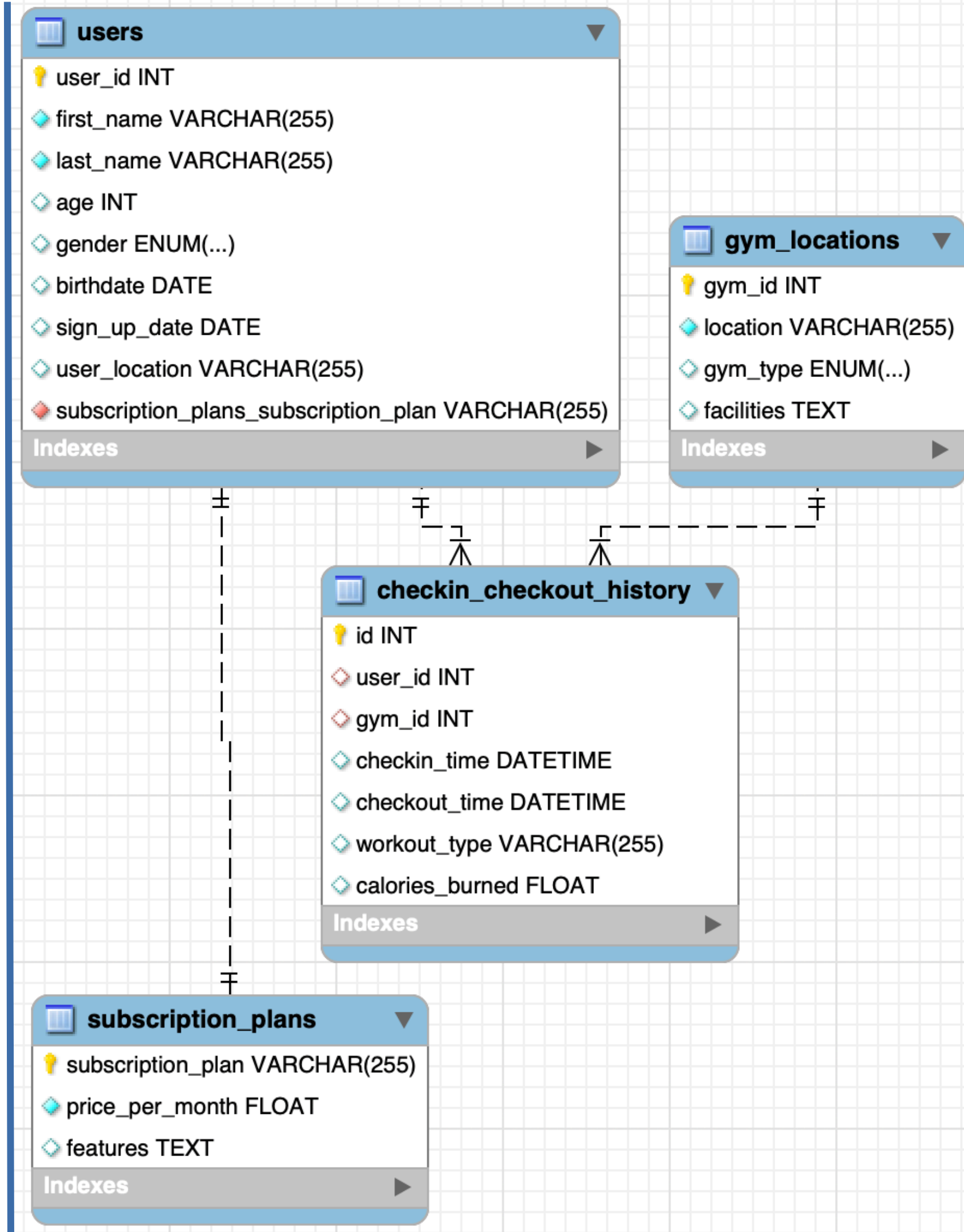
Data Manipulations
on KNIME workflow



Visualizations

ERR Diagram of the 1st Dataset

- Relational data on Gym Check-ins and User Metadata: CSV from Kaggle



The 2nd Dataset

- Census data : CSV from government database

City	State	Median Age	Male Population	Female Population	Total Population	Number of Veterans	Foreign-born	Average Household Size	State Code
Berkeley	California	32.5	60142	60829	120971	3736	25000	2.35	CA
New Britain	Connecticut	33.4	37350	35459	72809	2219	15080	2.52	CT
Alexandria	Virginia	36.6	74989	78522	153511	10635	44030	2.2	VA
Overland Park	Kansas	38.2	93355	93156	186511	10461	21407	2.41	KS
Melbourne	Florida	43.4	39180	40956	80136	8363	9685	2.37	FL
Sioux Falls	South Dakota	34.6	86596	84934	171530	8658	12934	2.38	SD
Huntsville	Alabama	38.1	91764	97350	189114	16637	12691	2.18	AL
Turlock	California	36.2	33190	39103	72293	3397	14686	2.76	CA
Gainesville	Florida	26.0	60803	69330	130133	4788	15272	2.33	FL
Champaign	Illinois	28.7	43326	42760	86086	3734	12261	2.25	IL
Inglewood	California	34.9	52995	58661	111656	2703	34171	3.12	CA
Saint Joseph	Missouri	35.7	37688	38408	76096	5846	3755	2.58	MO
Brockton	Massachusetts	35.2	46273	49041	95314	3036	27313	2.88	MA
Tulsa	Oklahoma	35.0	197437	205654	403091	24672	43751	2.37	OK

API Access



Variables

Average temperature during the day, moon phase, weather description, date

Location/Date Range

New York City

Jan 1, 2023 to Oct 15, 2023

Historical or Past Weather API

The Local Historical or Past Weather API (also known as City and Town Historical Weather API) allows you to access weather conditions from 1st July 2008 up until the present time. The API returns weather elements such as temperature, precipitation (rainfall), weather description, weather icon and wind speed.



1st July 2008 onwards




Astronomy

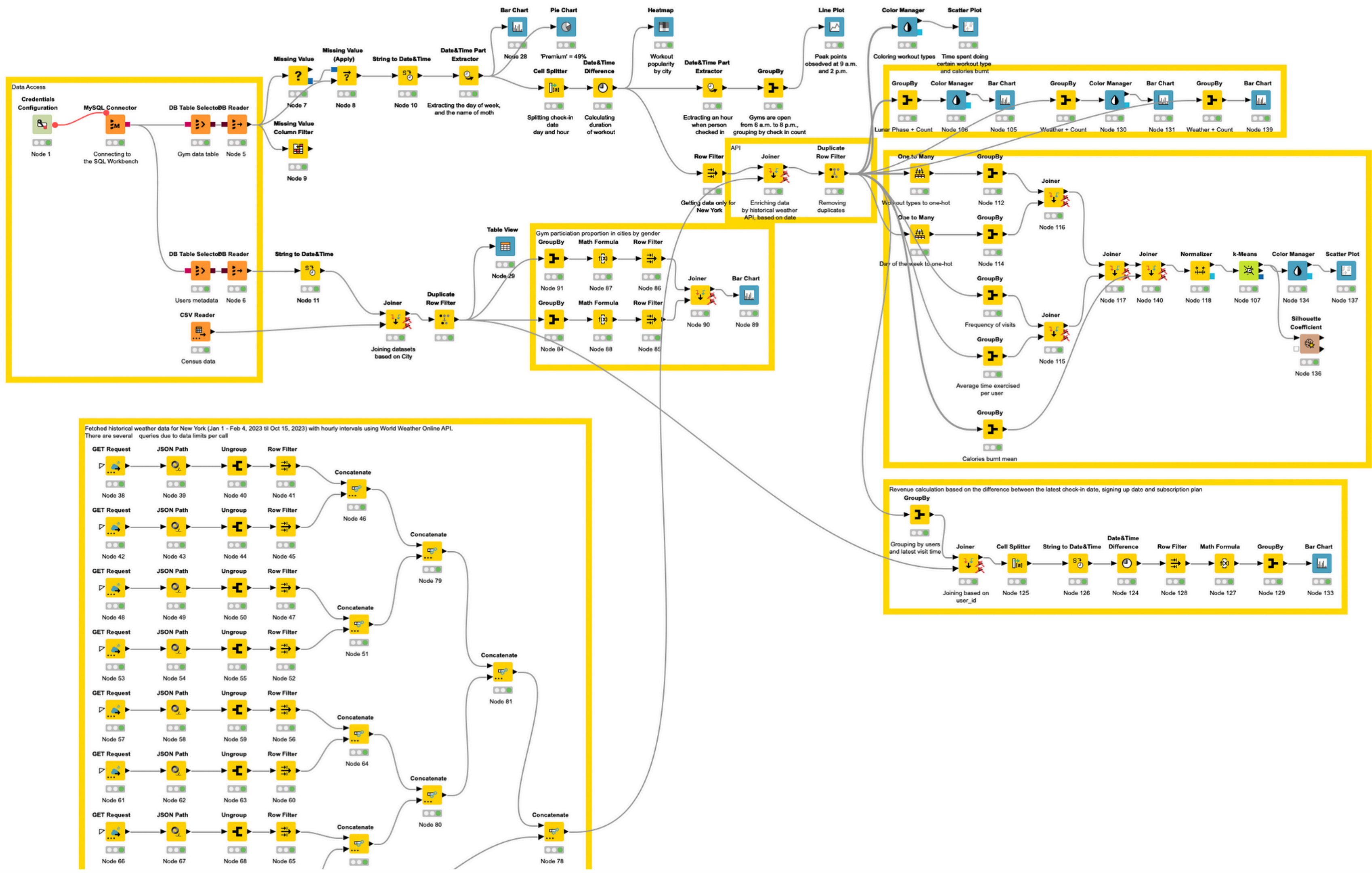


Hourly and 3 hourly intervals

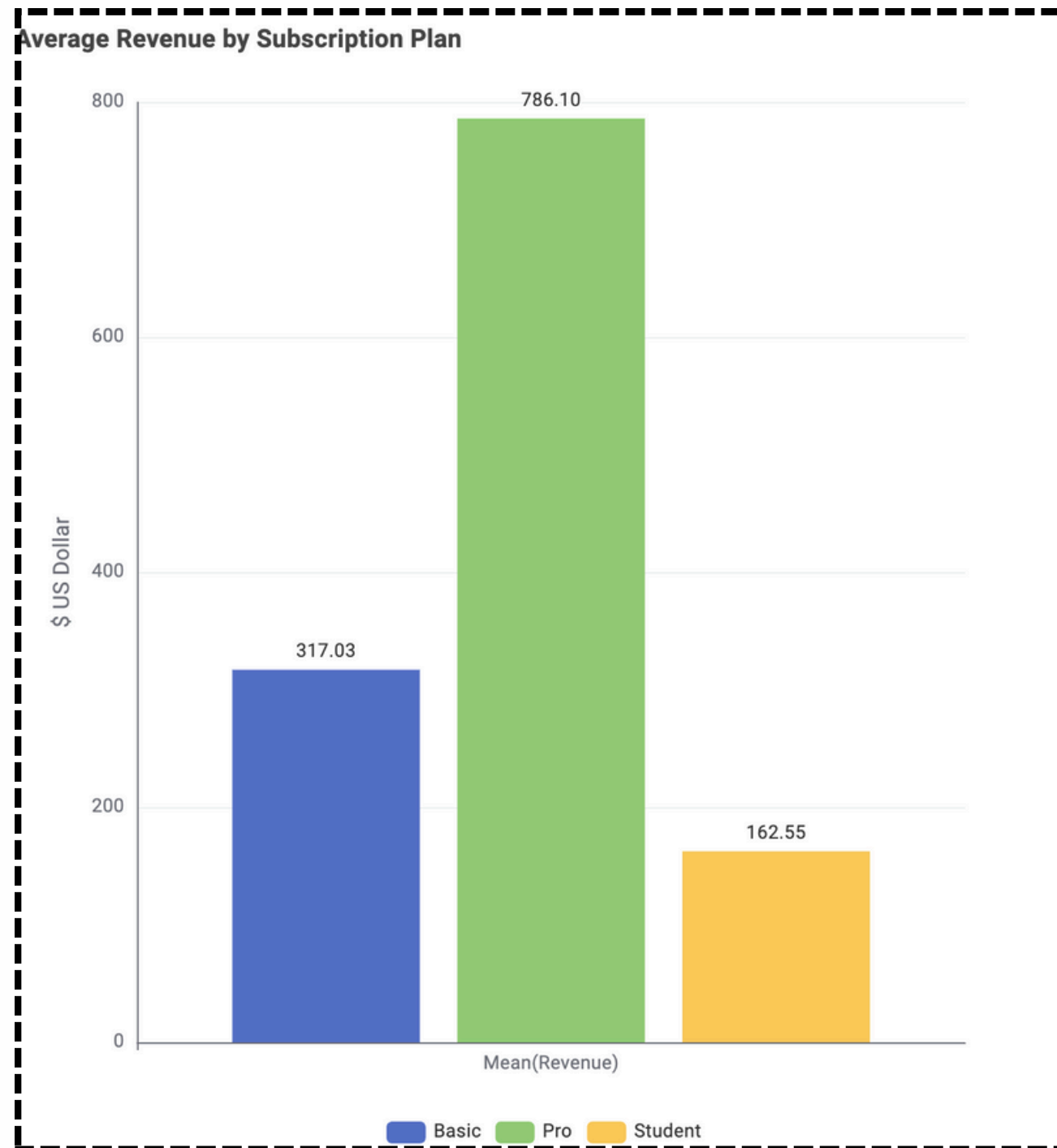


JSON, XML, JSON-P format





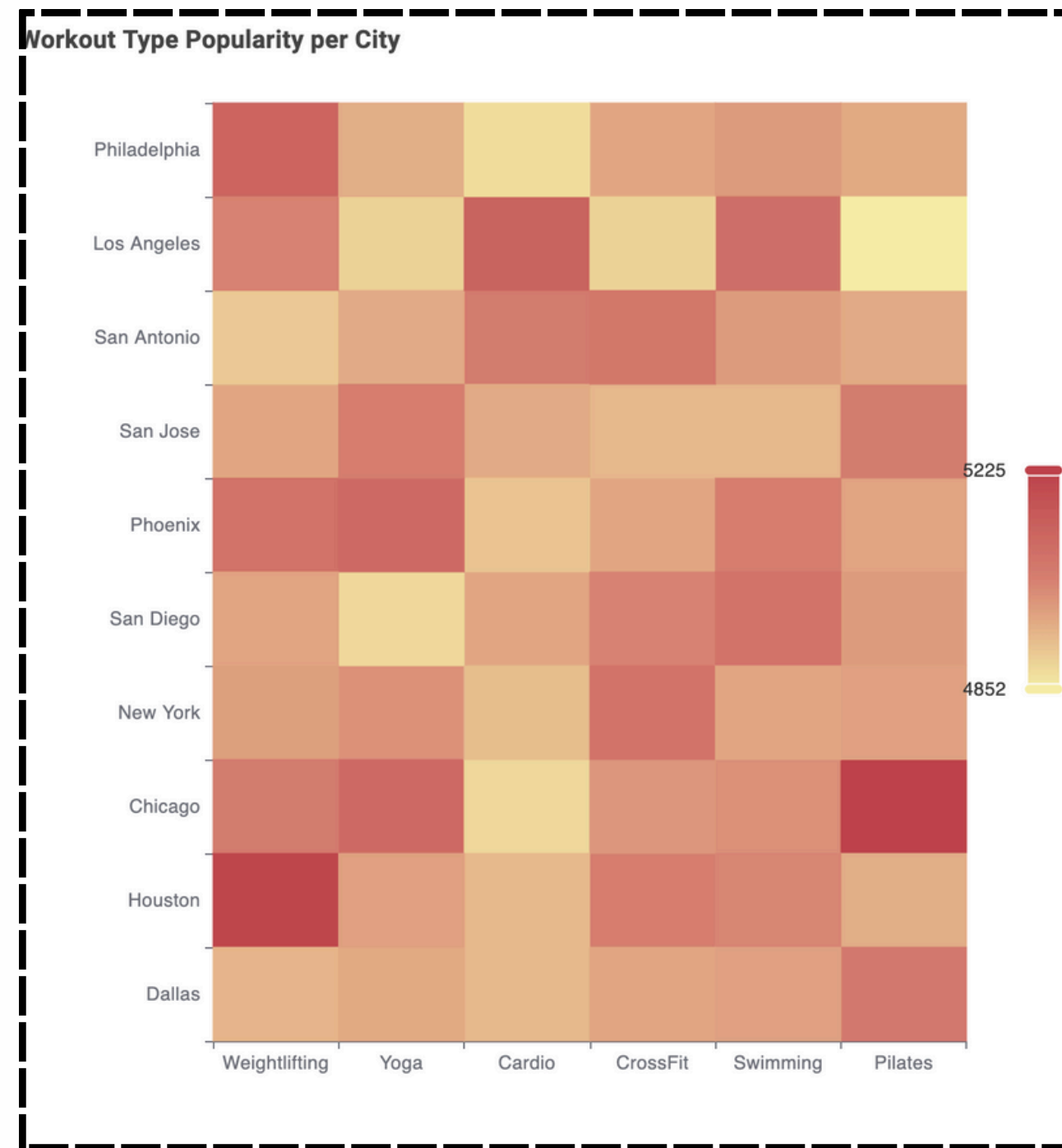
Revenues by programme



This graph represent the average dollar amount spent per user and aggregated it by the subscription type per number of months subscribed :

- Most of the revenue earned by gyms in the US comes from 'Pro' programmes;
- The 'Student' programme is the least profitable

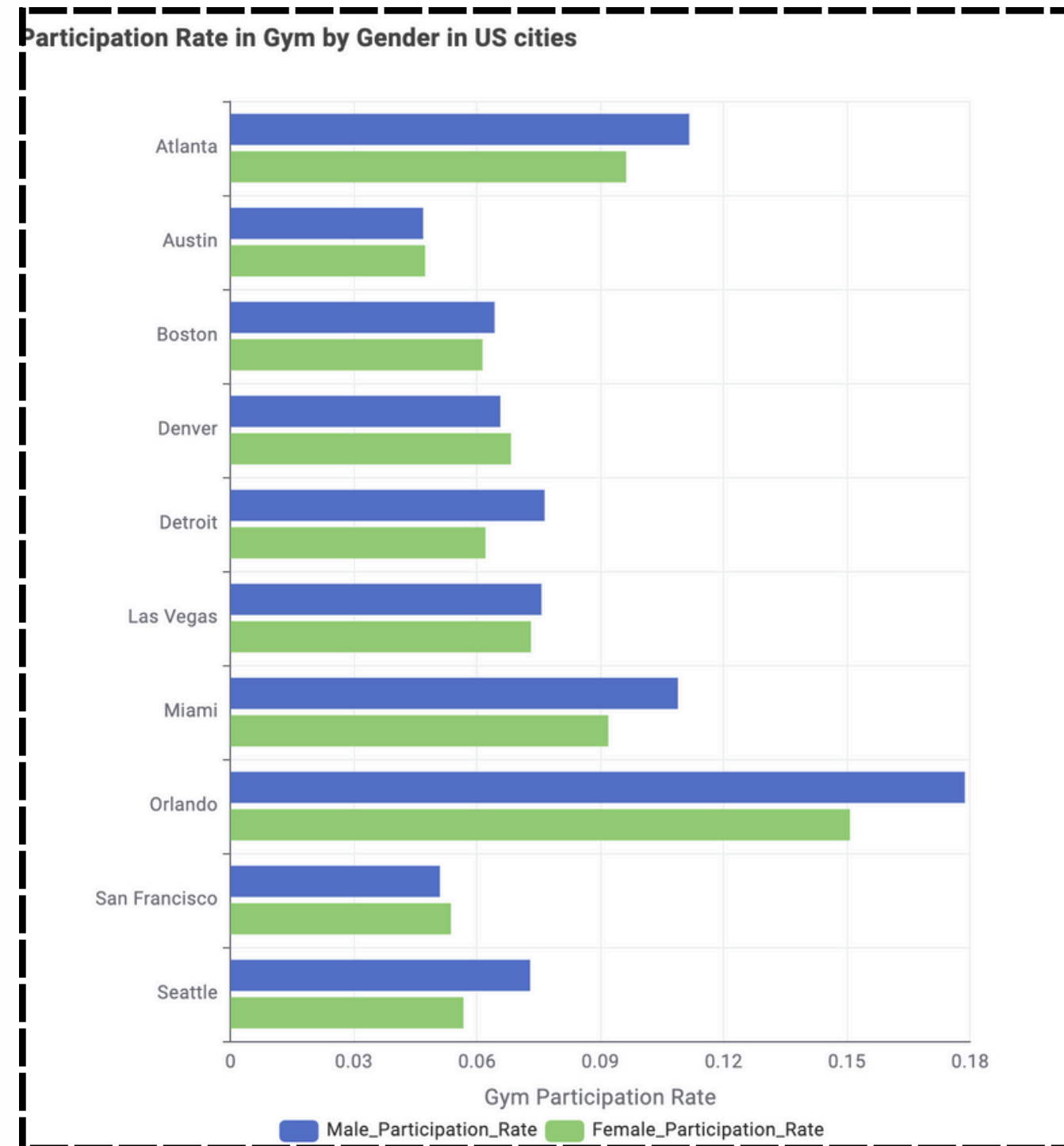
Preferred type of sport



This graph shows the gym participation rate by different cities' population by gender :

- Popularity of sports varies depending on the city
- Great interest in weightlifting in Houston and Philadelphia, cardio in Los Angeles and pilates in Chicago

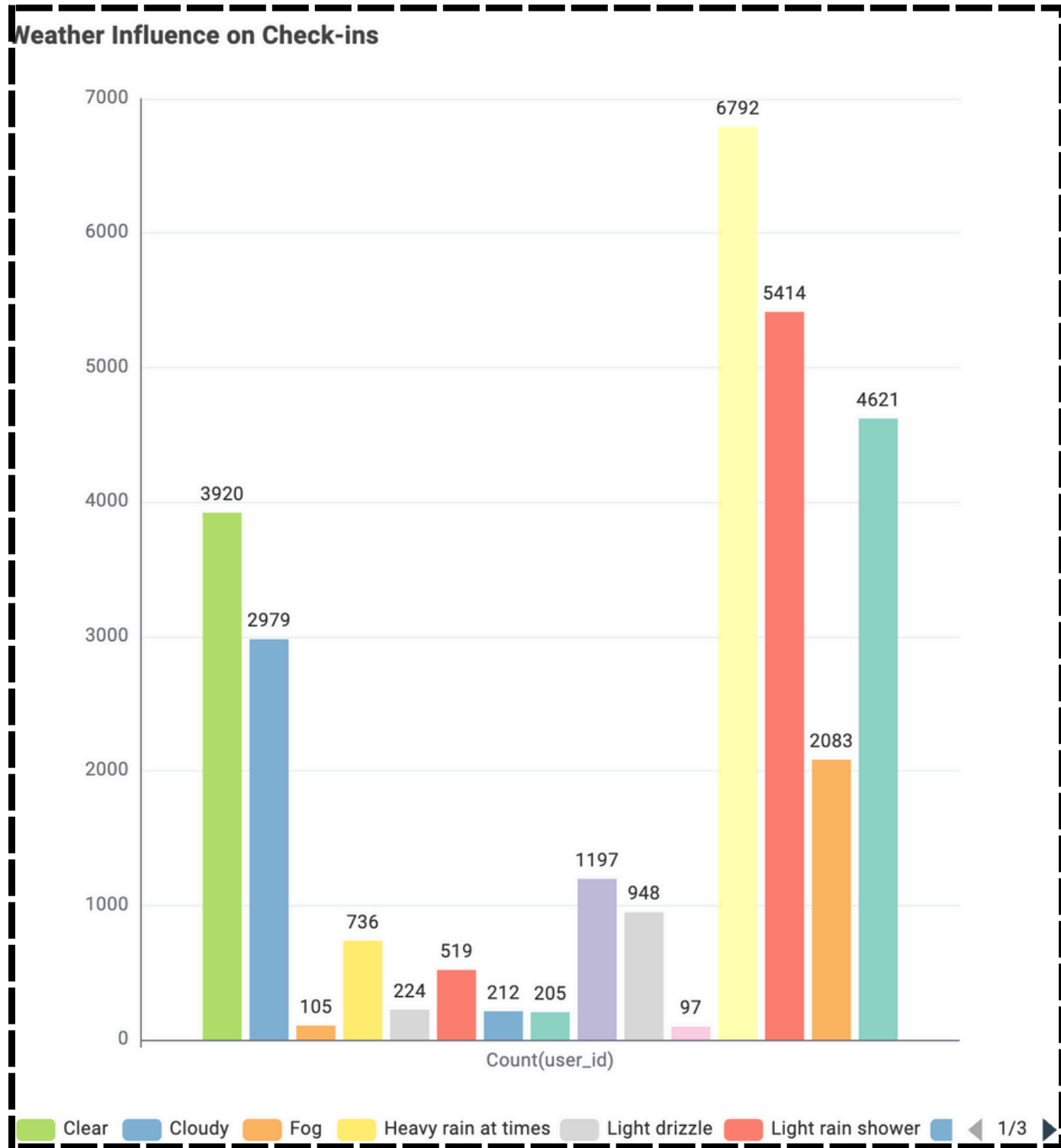
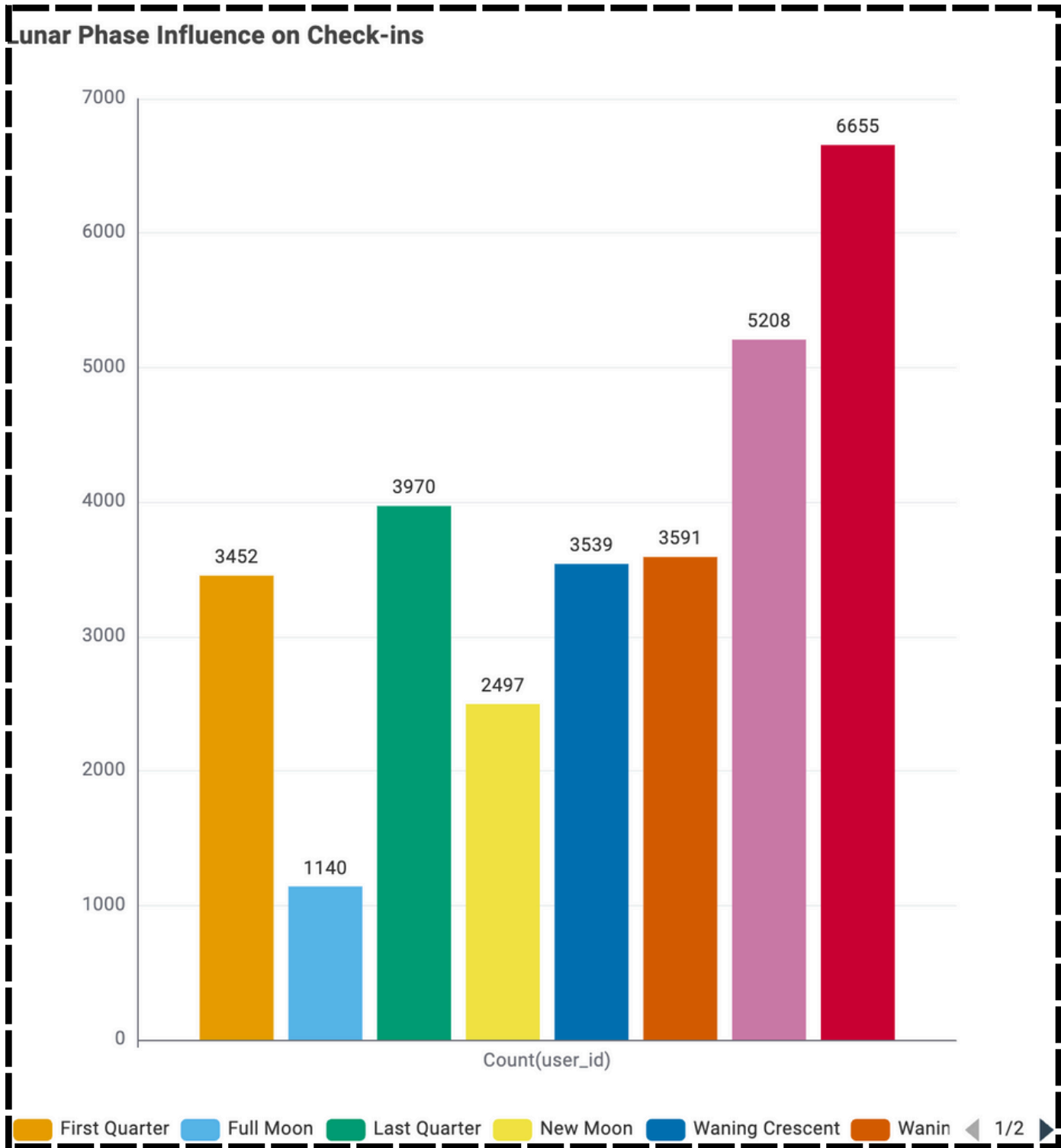
Preferred type of sport



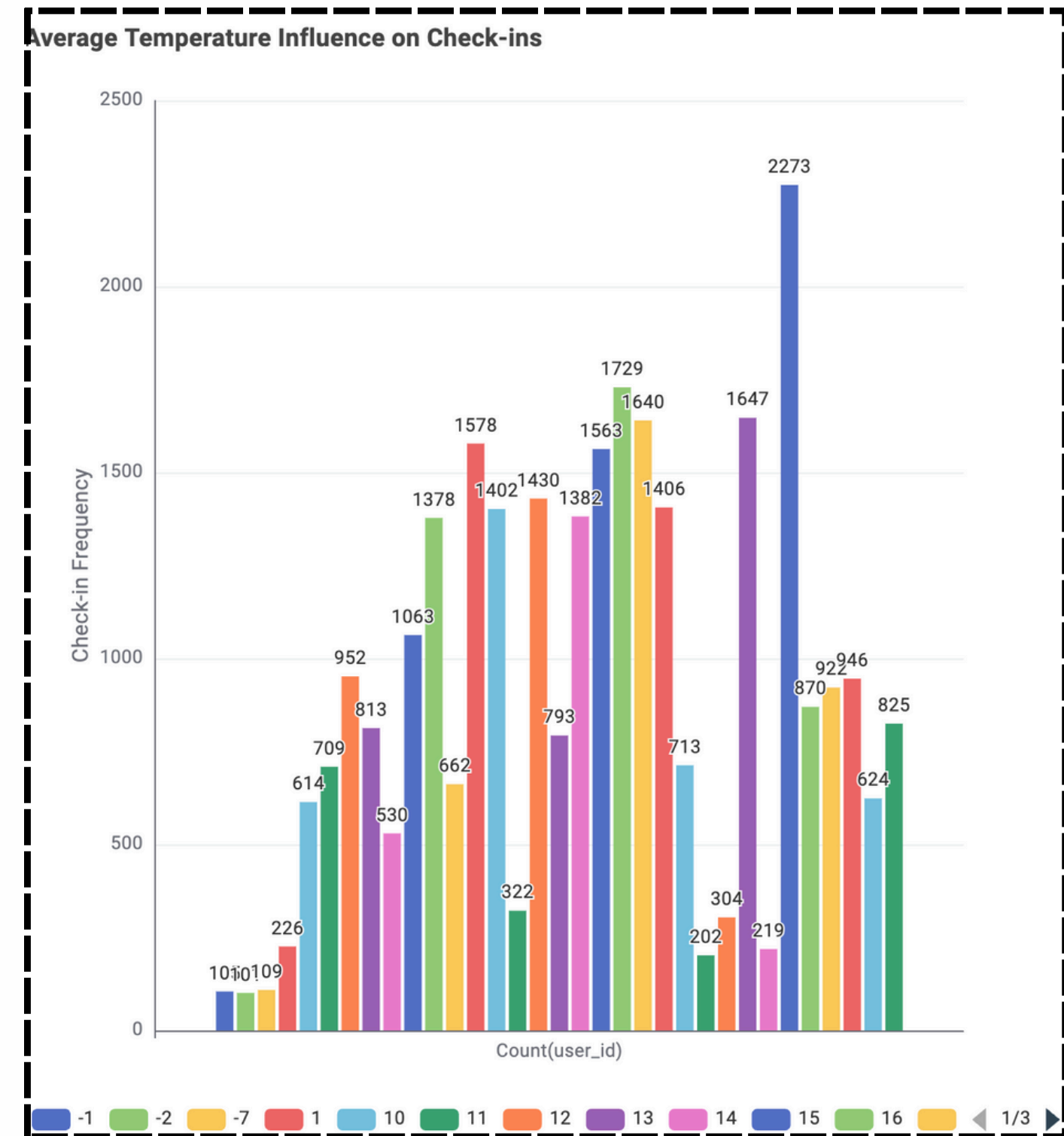
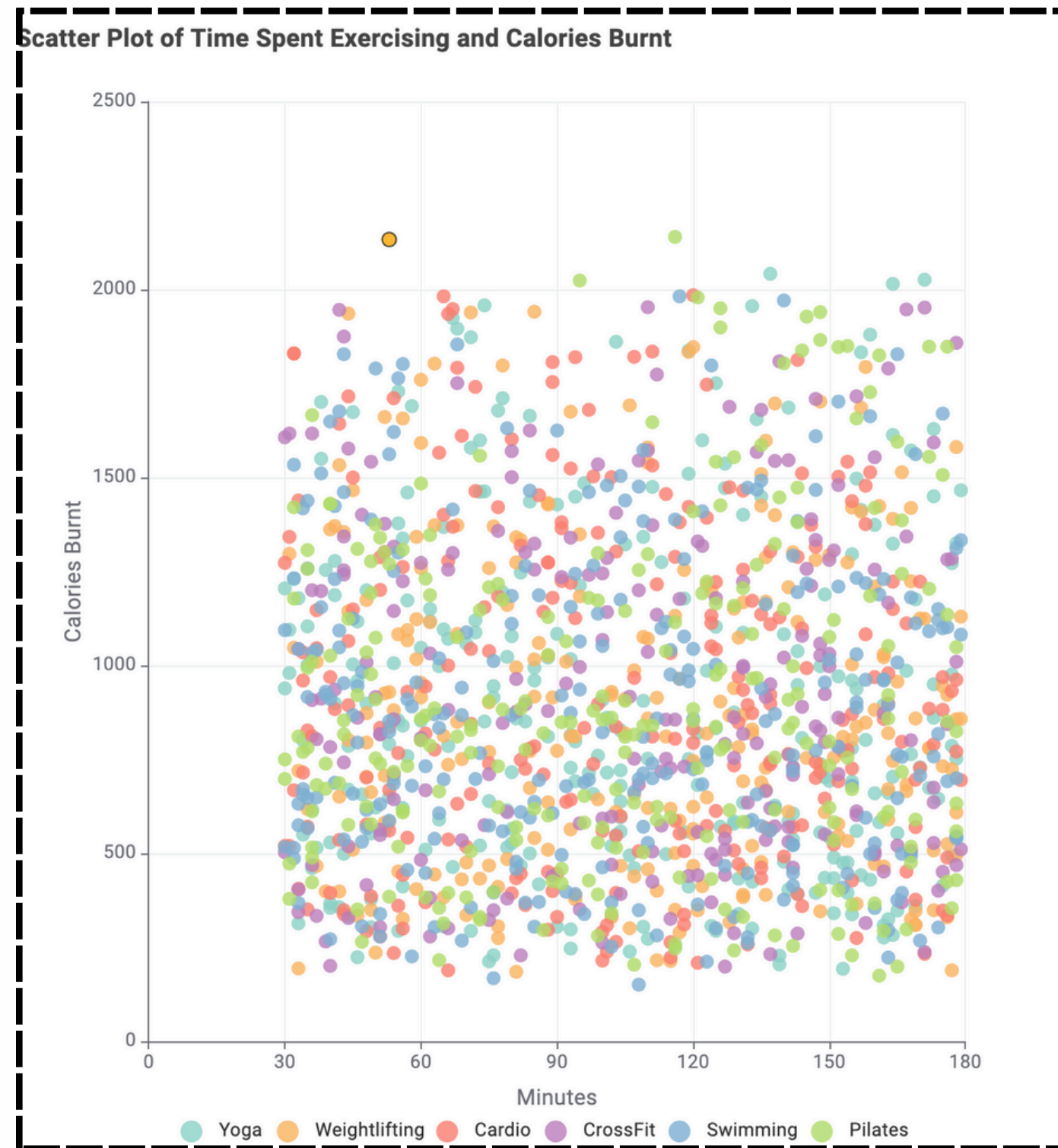
This graph shows the gym participation rate by different cities' population by gender :

- More men than women take out gym memberships.
- The difference is significant in Atlanta, Orlando, Miami and Seattle.

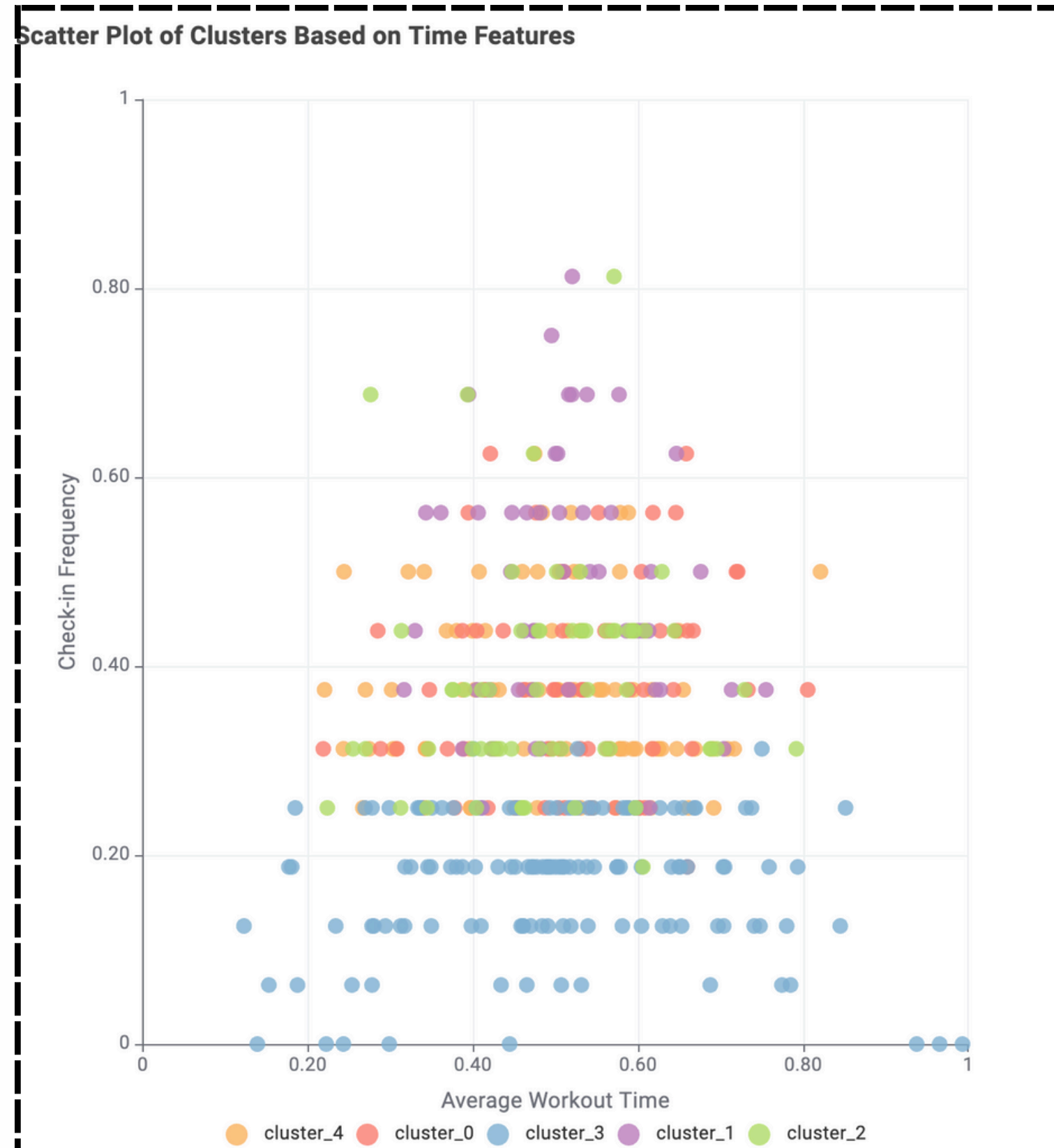
Analysis of external factors -New York



Analysis of external factors -New York



New York: K-Means Clustering



Conclusion



User behaviour varies according to their demographic characteristics and exogenous factors. In order to increase their income, gyms need to adapt their marketing strategy according to their location and the profile of their users, particularly those who subscribe to the ‘Pro’ programme.





Q&A

