Summary

This assignment implements three different analysis based on the checkins_melbourne dataset.

Part 1

Motivation and Problem

Melbourne is a very beautiful city, many tourists come here every year, so I guess most of the users who check in in Melbourne should be tourists. In order to verify my conjecture, I queried out the location of each user's first check-in and the last check-in location in chronological order, then respectively found the two areas with the most concentrated check-in locations through clustering algorithms.

Result

My analysis in Part 1 shows that city area and Melbourne airport area are the top 2 popular places with the most concentrated first check-in and last check-in. This result verified my assumption.

Part 2

Motivation and Problem

In the part 1, I have found that most of users in checkins_melbourne are tourists, so I've got interests on the change on the number of check-ins in different months and time. Meanwhile, this article states that Victoria's international tourists in 2010 increased by 7.7% compared to 2009. Therefore, I performed an analysis to research the pattern of the change.

Result

The results show that September 2010 has the most check-in number (2193), and compared to the year 2019, the number of check-in significantly increased. This may due to the fact that more international tourists came to Victoria in 2010. Besides, users prefer to check-in during the night time (after 18:00pm), and the check-in activity occurred the most frequently during the period from 24:00pm to 6:00am (i.e. Early morning).

Part 3

Motivation and Problem

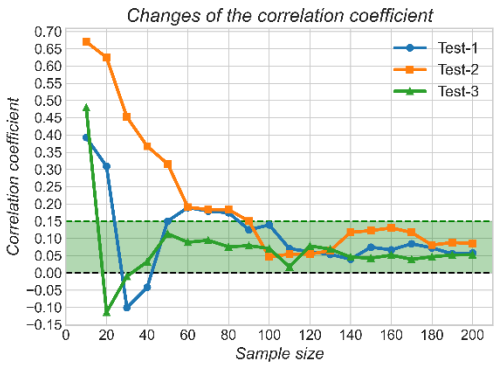
In daily life, people's behaviours may be influenced by friends. Therefore, I want to check whether the friend network influences the location of users' check-in. In this part, I transfer the friends dataset into an edges table for network query (which I think is the **most important part**), then I write a function to compute the friend network distance and spatial distance between check-ins and the correlation coefficient between them (which I think is the **most challenging part**). Since the different sample size will influence the correlation coefficient, I keep increasing the number of samples to observe changes of the correlation coefficient. Then I export the queried data into python to conduct the hypothesis testing (t-test) to get the significance (which I think is the **most original part**). Due to the limitation of running time, I can't use all the data for calculation, so I designed three experiments, each time randomly selecting different data for analysis.

Result

The figure shows the correlation coefficient which computed by pgsql, we can find that with the increase of experimental samples, the correlation coefficient of the three tests finally stabilized in the range of 0-0.15. Therefore, there is a weak positive correlation between the friend network distance and spatial distance of check-ins.

The table shows the result of t-test which computed by python, we can see the significance of three tests all greater than 0.05, which indicates that friend network distance and spatial distance between check-ins are correlated.

In short, my results show that the more familiar users are (i.e. the closer their friends are to the network), the closer their check-in distance will be.



Test no.	*Sample size*	*Corr.*	*Significance*
1	200	0.057	0.418
2	200	0.085	0.227
3	200	0.052	0.372

Tab. Result of t-test