CONTENTS

EXECUTIVE SUMMARY

This project aims to provide understanding in transport behaviours of Opal card users of Transport for New South Wales' public transport network. Through the procedure of this project, a background on the topic is reviewed so that the basis of research will be identified; the review of the background clarifies the aim for the project, which is required to analyse the mode choice, mode shift as well as spatio-temporal pattern of public transport users in order to come up with what truly affect their travel behaviours. Then, qualitative and quantitative research design will be utilised so that both the opinions of the transport users themselves as well as insights from their Opal data will be recognised for further analysis for transport pattern. The result from the project shows that transport pattern is decided based on the habit of travelling of the users, the available transport mode that is offered in the stations, the effect of off-peak/peak hour application, the accessibility to the station, and the infrastructure and service quality that the station can offer to its passengers. Therefore, immediate actions need to be done in constructing or enhancing the accessibility to public transport stations so that more passengers will become interested in the value that the public transport network can increasingly offer to them. However, the analysis of this project has just investigated only one side of a coin; further research on the impact of the significance of off-peak or peak hours to the schedule of timetable of the public transport network, or systematic procedure to improve data quality acquired from transport activities.

INTRODUCTION

Public transport (PT) is the easily accessible, common, and economic means of transport which is used by general public to commute. It is operated by following the scheduled time frames as well as on the predefined routes and every customer pay the allocated travel fares for their journey. For instance, trains, city buses, light rail, trolley busses and rapid transit (underground/metro/subway) are common examples of PT. Various airlines, intercity rails and coaches are providing the PT service for their customers and even high-speed rails are providing their services in some parts of the world. The most important business actor is PT network is the passengers as they are the main users of the infrastructure of PT. However, in order to attract new passengers as well as satisfy the needs of current passengers, it is important to investigate the transport behaviour of the passengers to understand what factors can have an impact on their travel patterns. However, without the correct approach to analyse the data acquired from PT activities, it is very difficult to understand the true meaning behind meaningless numbers. Generally, there are 4 factors that affect the transport behaviour as analysed in Literature Review: cost and accessibility, the users, the infrastructure, and the business and technical factors. Each of them has their own way to impact the transport behaviour of PT users, so investigating the spatio-temporal data, transport mode choice as well as mode shift of PT users may give the PT planners a more detailed view in the operation of PT network. Therefore, this project is initiated to study the de-identified data of Opal card users to understand the impact of mode shift, mode choice, as well as infrastructure accessibility... to understand the purpose as well as habit of PT users.

CHAPTER 1
BACKGROUND

The debut of NSW first means of PT, the railways, happened during mid-1860s, or 1855 to be exact, with the establishment of the first passenger railway line from Sydney to "Paramatta Junction", which is Granville Railway Station today (TfNSW n.d.). According to Robert (2009, p.3), this appearance was established in order to serve the pastoral interest as well as settlement for the migrants in search of gold during the gold rush era. As such, the later expansion of the railway system was made to serve the establishment of trade route from Sydney to other civilised areas. It can also be seen that industrialisation had led to the urbanisation in Sydney, especially during 1902, and railway system in NSW had been expanding since then. However, the most outstanding progress seen during the booming era of NSW railway history was during the First World War and the isolation during 1920s when NSW railway system kept expanding at rapid rate in order to satisfy all domestic needs in Australia through large-scale operations, which is depicted as a "huge vertically integrated empire". Nowadays, PT in NSW does consist of not only railway but also bus, ferry, and light rail, which make the PT system into a complex system. With the population of 7.6 million by September 2015 (BTS 2016), each citizen made approximately an average of 79 journeys on all transport modes for travelling every year. As such, it can be seen that the development of NSW has big impact on the expansion of PT network. However, the citizens are always at odds with the government over the traffic development. As reported by Andrew (2018), recent protest against the expansion of motorway network had become a group rally where opposition groups allied in an effort to bring the issue of PT to the cabinet. It is reported that people suddenly got hit with a toll charge on the previously free public road they were travelling, so the action was quickly taken, and it soon became a rally to ask for rectification of NSW PT. Thus, the opinions of the citizens to the government are that they are not satisfied with the current state of the PT and demand the government to upgrade or change it soon instead of focusing on enlarging motorway network and collecting more toll charges. As analysed in Literature Review part, the factors that affect the citizens' choices as well as travel behaviours related to PT are categorised into 4 groups: cost and accessibility, the users, the infrastructure, the business and technical; therefore, in order to understand what lies behind people choice of choosing PT, more research should be conducted to investigate those factors.

Generally, PT vehicles follow the fixed routes in order to provide their services with pre-planned timetables and they mostly follow the 15-minute-interval between two services but most of the trips may include the other modes of travel, for example, PT users may walk or use bus services to reach to the train station (McLeod, Scheurer and Curtis 2017). Shared taxis also provide the on-demand facilities in many parts of the world and they contribute to bring the passengers to the interchanges. Paratransit is also in use for providing door-to-door service, but its demand is less than other types of services. Additionally, urban public transport may differ between different locations such as Europe, Asia, North America... In Asia, private-owned, publicly traded and profit-oriented PT is in use (Calimente 2012); in North America, mass transit operations are conducted by municipal transit authorities; in Europe, both private companies and state-owned companies operate the PT services. Basically, two types of travel fares mode may be implemented, it could be either pay-by-distance or funded by government subsidies in the form of flat rate fares; services could be profit-oriented through high farebox recovery ratio, high usership number or it could be regulated and probably

subsidised from national or local tax revenue; some PT services are subsidized and completely free of charge. So, it can be seen that some components such as historical, geographical, as well as economic reasons could be the fact that differs the nature and property of PT services. Basically, Old World countries are conducting frequent and extensive systems to serve their dense and old cities; but at the same time, New World countries are implementing less comprehensive and more sprawl type of PT system. According to Mars et al. (2016), some of the studies related to PT are based upon analysis of attitudes and perceptions, impact of urban environment or social connections on travel behaviour. Qualitative study is the well-known approach of analysis in the fields such as Sociology and Psychology. Such kind of study is increasingly implemented in Transportation Planning and Engineering which deals with study on road safety and PT service quality. Thus, it can be seen that qualitative study has been playing a significant role in investigating the travel behaviour of commuters, which is early depicted as in the workshop of Grosvenor (2000) in which it is claimed that qualitative research plays a key role in transport sector policy and research whether it is conducted independently or in association with quantitative method, but if it is used sensitively with complement of quantitative research, human behaviour can be fundamentally studied in terms of feeling and emotion; or in the work of Kelly and Study (2003) where qualitative research is deemed as a tool to solve the complexities of travel behaviour to unravel the hidden problems as well as forecast the impact of them to the trends in the future since human behaviour towards transportation may not be rational as usually defined rational behaviours. As well, the factor of users itself can be uncovered by conducting both qualitative and quantitative research as one of the reviewed literatures suggests that the travel patterns, in Riga agglomeration particularly, can be discovered via analysis of spatial patterns of youth travel, motives and factors influencing transportation choice for school trips and travel times (Girts 2014). These literatures contribute to the fact that both qualitative and quantitative research can be used to discover travel behaviours of the PT users through their responses or past PT usage records, such as their transport mode (including single- and multi-mode), transport line or patronage based on recognition of their spatio-temporal travel records.

At the outset, many people use private vehicles for their commuting purpose, and hence there is the necessity of appropriate policy which could minimize the dependency of people on their private vehicles and encourage them to think about an alternative which is the use of PT. Such type of policies should be able to improve the services of public transport and encourage people to utilize those services. There is also a necessity of promoting measures which could minimize the private vehicles including car (Tommy and Geertje 2007). On the other hand, Sandiv and Genevieve (2015) provides an empirical research on transit patronage determined by service reliability which provides that reliability of the service in fact has an impact on the transport patronage, transit mode choice as well as route selection. Besides, increasing the use of PT and minimizing the uses of private vehicles are very significant but more challenging tasks, especially while dealing with urban transport. There is the necessity of PT to be more competitive and market-oriented since they are also considered as a service product. During the past decades, many PT services have been privatised in which the financial performance of private service providers is the key component. The loyalty of public transit passengers is very important since they can play vital roles for the long-term financial performance and thus it is a primary source for the competitive benefit. For the manager of public transit, especially for the newly operated public transit, understanding of travellers' behaviour is

very important because such type of information could be helpful for creating appropriate strategies for meeting travellers' requirements; hence, it could help to satisfy the regular customers and can attract new customers who are using other types of transport modes (Wen-Tai and Ching-Fu 2011). Incidentally, the Smart Card Automated Fare Collection (SCAFC) system (a.k.a. Opal card system) has become an effective tool in PT and been helping to collect huge amount of data every day in currently running system. Although, the primary purpose of using SCAFC system is to collect revenue, but on the other hand, it is helping to collect more data from the travellers; and by analysing those data, it is possible to understand the travel pattern of commuters and hence help to develop different strategies and plans in order to fulfil the users' requirements. In fact, it can collect the users' data related to exact travelling time, location, numbers of commuters and many more. Currently, those datasets have increased rapidly so there is the necessity of sophisticated tool having capability to retrieve, mining and analyse them so that the result can provide some relevant information and by following that information the service quality of PT could be improved (Bruno, Catherine and Martin 2006). Thus, the first aim for this report is to investigate the mode choice or mode density of PT through the spatio-temporal data as well as survey data collected from PT users to analyse the impact of mode choice and mode density to the behaviour of PT users.

Another aspect that needs investigating is the cost and accessibility of PT mode as one of the most significant factors which may affect for using PT is the access time or distance of station, terminal or stop. Basically, the transport prices include travel time, monetary (money) costs, risk and discomfort. Prices have some effects on consumption, which means that price changing has a great influence for the consumption of goods. Considering the law of demand, which states that when price reduces then consumption increases, and when the price increases then consumption decreases, transport activities also follow this format; so, when travel fares increase, the mobility or usage of transport declines, and when travel fares decrease then mobility of the commuters increases. The variation in travel fares could affect the route, trip frequency, destination, mode, vehicle type and schedule, type of service selection and location of parking (Todd 2017). As identified by Becky, Cynthia and Eric (2010), the older the existence of a railway line is, the higher patronage rate they have; as well, stations with interchanges produce higher patronage than others without interchanges and the factor of inter-modal transfer is very important since it affect the patronage of the PT station itself. Also, per Yulin and Phil (2013), the roll-out of smart card technology may lead to the successful introduction for peak/off-peak fare; as such, the introduction of Opal card has opened the possibility for PT planner to implement this concept into NSW PT network. As can be seen from the status quo of the PT network, people travelling using train mode during 7 AM to 9 AM in the morning and 4 PM to 6:30 PM in the afternoon have to pay fare 30% higher than travelling outside of this time frame. However, this implementation has not solved the problem of peak spreading where successful implementation of policy may lead to less concentrated activities during peak hour. Also, as mentioned by Yulin and Phil (2013), solving peak spreading by implementing differential fare is only a short-term solution while long-term solution may involve capacity increase for the PT network. As such, the second aim for this report is to investigate how passengers react to the implementation of peak/off-peak hour by analysing the density and the trend of tap-on and -off event occurring during the day to discover how the cost affects the passengers as well as how the accessibility to the PT station influence the density of patronage across NSW PT stations.

On another matter, the effect of infrastructure to the transport behaviours is very transparent since the infrastructure implies about huge construction with complex transport system (both IT and business) fitted in to serve the demand of its passengers. As a matter of fact, infrastructure plays the vital role for providing the quality service to commuters. In order to obtain this goal, the rail service should be able to expand their network coverage area, provide easy access to the station as well as at the interchange point, increase the service reliability and minimize the travel time. Some of the factors such as multi-modal transport, accessibility to the train stations, distance to the station and the inconvenience while changing the trains during journey indicate that it is a primary factor which determines the level of service quality provided by train service (Martijin, Moshe and Piet 2009). Also, transportation is the key component for the energy shortage and environmental pollution so; transit-oriented development (TOD) has become a very useful strategy for reducing these issues. The main objective of TOD is to encourage people to use PT instead of private vehicles, to live near the stations or stops and attracting them to use cycle or walk for services, work and shop. This type of strategy is implemented by policy makers or planners. It is concluded as diverse, dense, pedestrian friendly land uses near transit nodes that, under the appropriate conditions and convert into higher patronage. So, increasing patronage is the key concern of many transport agencies. For example, one of TOD methods, integrated land use-transportation planning, is in use for more than half a century. While applying TOD, it is necessary to consider about the local culture, economy and history. In fact, TOD is a planning strategy which is focused on uses of mixed land near the public transport stations by ensuring safe access for cyclists and pedestrians. There is no particular definition of TOD because of its varying nature based on historical and geographical conditions and also it could be change after long time and depending on the location (Li, Nicholas and Michael 2011). As such, the third objective of this report is about how the infrastructure of the stations in NSW PT network affects the transport behaviours of its passengers.

Lastly, it can be seen that the amenities of qualitative PT services must build a major consideration of PT operators, for example, when crisis arises then its goal becomes more significant than ever. Actually, crisis generates the opportunity to encourage new users/customers, with who to develop long-lasting relationship to be established when the crisis ends. For the purpose of increasing user satisfaction, operators must focus for the persistent improvement of the facilities. For instance, due to the result of crisis in Greece, PT operators were merged in Athens, ticket sales outlets were terminated, ticket types and prices have altered and there was severe decrease in income of both drivers and users. This is a representative result of crisis and such type of incident could affect the customers' satisfaction and could minimize the demand of public transport (Dimitrios and Constantinos 2017). Besides, Intelligent Transport Systems (ITS) has been implementing with various initiatives during last two decades. In case of Europe, the "ITS Action Plan" finds many applications as primary components contributing to the effective co-ordination of transport chain. The experience and context around the current popular development of technological devices and Information and Communications Technologies (ICT) platform bears the exposure of ITS are important in which they permeate the logistics chain and the transport (John and Corrine 2013). In Sydney, Opal card (smart card ticketing system) was implemented in a phased basis from 2013 to late 2014 which later replaced the existing paper-based ticketing system and stood up as the main ticketing system for the public transport. In fact, Opal card is basically similar to the other systems which are in use in PT all over the globe such as Hong Kong's Octopus card and London's Oyster card

and it has made PT journey very easy, simple and intended to increase more convenience, but the sole difference is in the travel fare structure. The improvement in convenience was made possible due to the use of Opal card instead of using paper-based ticketing system (Richard et al. 2017). Hence, the last objective of this report is that it examines the business as well as technical perspective of NSW PT network through the use of de-identified Opal card data so as to discover the needs for change or improvement if the current PT network has not fully satisfied the conditions that PT users expect.

CHAPTER 2
METHODOLOGY

The research conducted in this report follows the design of a mixed research methodology where the research team utilises both kinds of research methodology: qualitative research and quantitative research. Inspired by the analysis of Shannon and Christina (2018), the design for mixed research methodology is chosen as embedded type, in which the research team carrying out this research methodology must have two sets of data, one for qualitative and one for quantitative, with one set supporting the other. The purpose of this choice is that the dataset given by the research team supervisor is from 2016 and there have been so many changes in the public transport system since then; so, the given dataset may not reflect the status quo of the system correctly. Using qualitative design to discover the opinions and behaviours of the transport users may help building up a basis for digging into the quantitative design and obtain reasonable results. Additionally, a research on the background of the research topic has been carried out in order to clarify the knowledge base from the Literature Review part. The focus of the background is on the history of the transport system, the problems that have been the controversy between the public transport planners (a.k.a. the government body of public transport system) and the transport users, the relationship between those problems and the factors analysed in the Literature Review part as well as the role of the research topic in this relationship. Thus, the structure of this report has been divided into 4 chapters: Background, Methodology (the current chapter), Research Design, and Discussion. In some chapters, there will be some integral parts, which can be seen in the Table of Contents.

By the result of researching the background in the previous chapter, the focus of the qualitative research will be on the following perspectives:

Perspective	Content	Inspired by
Users' opinions	The opinions of PT users on the PT can be discovered by acquiring their experience when using PT service during peak hour	Wen-Tai and Ching-Fu (2011)
Mode choice and purpose of choice	The behaviour of PT users when choosing transport mode relies on the perceived values that the transport mode can offer to them.	Sandiv and Genevieve (2015)
Infrastructure and service quality	The impact of infrastructure and service quality on PT users' overall transport experience may be high and it needs investigating.	Martijin, Moshe and Piet 2009

Table 1 - Inspiration for qualitative research acquired by reviewed literatures

From Table 1, it can be seen that the perspective presented may depend heavily on the bias of the respondents when conducting a qualitative survey since every person has different levels of perception and point of view, therefore, appropriate sample size and sample collection must be employed so that the result from qualitative research maybe useful for the quantitative research when combining them together.

On the other hand, the focus of quantitative research will be on the following perspectives:

Perspective	Content	Inspired by
Mode choice and time of departure	The PT users usually choose their PT mode as well as time based on their needs, no matter if the time they want to go is during peak hour or not.	Richard et al. (2017)
Mode shift effect	The ease of interchange may influence the travel behaviour of the PT users since they will find it more comfortable if it requires them less time transferring from transport mode to another.	Todd (2017)
Infrastructure and service quality	Infrastructure may hold the key to which attract the PT users since stations with sophisticated infrastructure can accommodate more transport needs than the one with less sophisticated infrastructure.	Martijin, Moshe and Piet (2009)
Crowding effect on busy stations	Crowding may have bad impact on the experience of PT users, but crowding does not occur in PT network all the time but during a specific time at a specific place. By uncovering the crowding effect on busy stations from when the crowding first starts and later reaches its climax, the patterns of PT users will be examined for future development of upgrading.	Alejandro, David and John (2013)

Table 2 - Inspiration for qualitative research acquired by reviewed literatures

As can be seen from Table 2, the method which the dataset is cleaned and processed must abide to the inspiration presented in this table in order to achieve the research objectives discussed in the Background chapter.

Hence, the mixed research will be conducted in the next chapter, in which their methodologies will be discussed base on the above inspirations as well as actual situation of the population or structure of the available datasets.

CHAPTER 3
RESEARCH DESIGN

A. QUALITATIVE DESIGN AND DATA ANALYSIS

I. INTRODUCTION

The qualitative research part of this research focuses on capturing the opinions of the customers in conjunction with their ages, occupations, frequent patterns, and experience when using the PT service. The method by which the information can be captured is through conducting survey. The channel which the questionnaires are spread is through online functionality; a Survey Planet account is created, and the questionnaire is implemented into the collection. The main focus of the questionnaire is from the 4 categorised factors that were discussed in the previous chapter. As stated by Kahle (1997), real-time responses from the survey can provide many advantages on to a focus group in order to rapidly assess the situation; they also provide timely, precisely, and probative information. In addition, as inspired by Dick (1998), an online survey carries a lot of benefits with it, such as controllable data quality, fast result, and cheap price. As such, conducting an online survey in order to collect PT users' points of view is a necessary and reasonable step to preliminarily investigate PT users' behaviours from reliable source of information; the result from the qualitative stage of the research will have a good impact to its quantitative stage since qualitative result will give a direction for analysing given dataset and explain some hidden meaning behind it.

II. METHODOLOGY

As per introduction, the survey used in qualitative stage will be conducted with an online service to collect the information from PT users. An online survey account was created on Survey Planet website to design the questionnaire structure as online survey can be conducted without location constraint. The target area is chosen as within Sydney and Wollongong area since Sydney is the largest city in NSW as well as Australia, and also the heart of PT system of NSW, whilst Wollongong is the third largest city in NSW (after Sydney and Newcastle) with a large PT network connected to Sydney. The combination of Wollongong and Sydney can cover the lack of information acquired from Newcastle since they accommodate large number of residents and are enough for a simple random sampling, the main method used in this qualitative research. Per William (1977, p. 18), simple random sampling carries the most convenient meaning that a proportion of sample is selected among the population so that every possible distinct samples have an equal chance to be drawn; so, this method is quite suitable for a large population size such as NSW population. Additionally, according to Robert (2015), random sampling can eliminate the influence of unpredictable or uncontrolled factors where respondents are chosen randomly from the population of people who are in inclusive criteria. As known that Sydney has the highest incidence of PT use among Australian capital cities, which was used by over 26% of residents for work or study (ABS 2008), it is ideal for random sampling since the chance that the respondents use PT daily is very high. Nonetheless, the sample size is a very important matter as it must be able to represent the whole target population. As per proof from William (1977, p. 76 &), as the population size is large (7.6 million), the sample size can be calculated using the following formula:

$$Sample\ size = \frac{z^2 \times p \times (1-p)}{e^2}$$

Where: z represents z-score

p represents probability of picking a PT user within population

e represents margin of error

At confidence level 95%, the z-score is determined at 1.96, margin of error is 5%; the probability is assumed at 0.26 (26%) as aforementioned; thus, the sample size needed for this survey is 296. Assuming that 90% of the returned questionnaires is valid, the required survey size is 329. The participants in the survey is determined as the people within the connections of research team members as well as their respective connections (e.g. family, friends, co-workers...). This selection utilised the web of connections created by social media to distribute the questionnaires to random people (i.e. friends of friends of friend...) without considering their status so that the questionnaire results will be unbiased and can be used accordingly.

After determining sample size for the survey, the most important step after that is designing the questionnaire. Initially, there are 3 questions that need answering when designing this qualitative research questionnaire:

- Question 1: What are the opinions of PT users? To what extent do they consider the services dedicated to the infrastructure regarding quality?
- Question 2: What passenger type uses PT the most? How often do they? What transport means do PT users ride the most?
- Question 3: What is the experience of PT users when using transport service, especially during peak hours? What is their overall rating over the service?

These questions are related to the factors that are analysed in previous chapter. Thus, the questionnaire is designed as per Appendix A, which is a collection of 12 questions in both open-ended and closed-ended types. In this questionnaire, question 4 is both closed-ended and open-ended question as the response varies depending on the respondent; questions 1, 2, 3, 5, 9, 10, 12 are used as closed-ended questions in order to limit the answer variety, but the scale is still within the context of the population; and questions 6, 7, 8, 11 are purely open-ended questions in order to capture the variety in responses as they represent the diversity of the research population; however, the level of response variety in the questions can be still considered generic and without bias. Among the aforementioned questions, questions from 1 to 8 are used to classify the respondents into different groups known as gender, age (according to the age standard clarified by ABS (2014)), economic status, frequency, purpose, usual departure time, and frequent destination whilst question 12 is used to measure the satisfaction level of the respondents towards the PT service; meanwhile, there must be a good reason for the responses used in question 9 and 10 as they have to cover the PT users' behaviours to some extent so that there will be no plausible responses to be missed. As per research by Craig, Brian and Jillian (2016), there are 3 variables that have very high correlation with perceived satisfaction of PT users: convenience, cabin environment, and ease of use. In addition, according to Lauren et al. (2013), the importance of reliability and frequency in PT service is, without a doubt, the core demand from PT users; nevertheless, it is advised that applying restrictive effect on private vehicle use does not have great impact to mode shift from private vehicle to PT but the accessibility to PT, comfort during travel, as well as pricing have most impact on mode shift behaviour. As such, the aspects that are to be evaluated in questions 9 and 10 are retrieved

from these elements to discover deeper into the attitudes of PT users towards the service they are using. As the last-mentioned question, question 11 carries a special meaning within its existence. According to NSW IPART (n.d.), Opal fares for passengers during peak hours are 30 percent higher than those during off-peak hours, which is a means to distribute the demand more evenly during the day as well as cover the increased marginal cost for service during this busy period. However, as report by Matt (2017), overcrowding situation has been rapidly increasing within the public transport system, especially Sydney's trains, when the demand is inclining in a day-by-day basis. This incidence has eluded the purpose of administering weekday peak and off-peak fares as the policy couldn't avoid the fact that the population of some big cities such as Sydney, Newcastle, Wollongong... has been increasing incessantly. Consequently, passengers will become more and more inconvenient with overcrowded PT vehicles they are riding in, and their concerns will be revealed within the context of this question.

On the other hand, the flow of the questionnaire is designed so that the respondents will be classified after the first 4 questions whether they are really PT users or not; as for PT users, they will follow down the structure of the questionnaire in a fashion that it will guide them through their procedure when choosing PT service. The flow of questionnaire can be summarised in the following diagram:

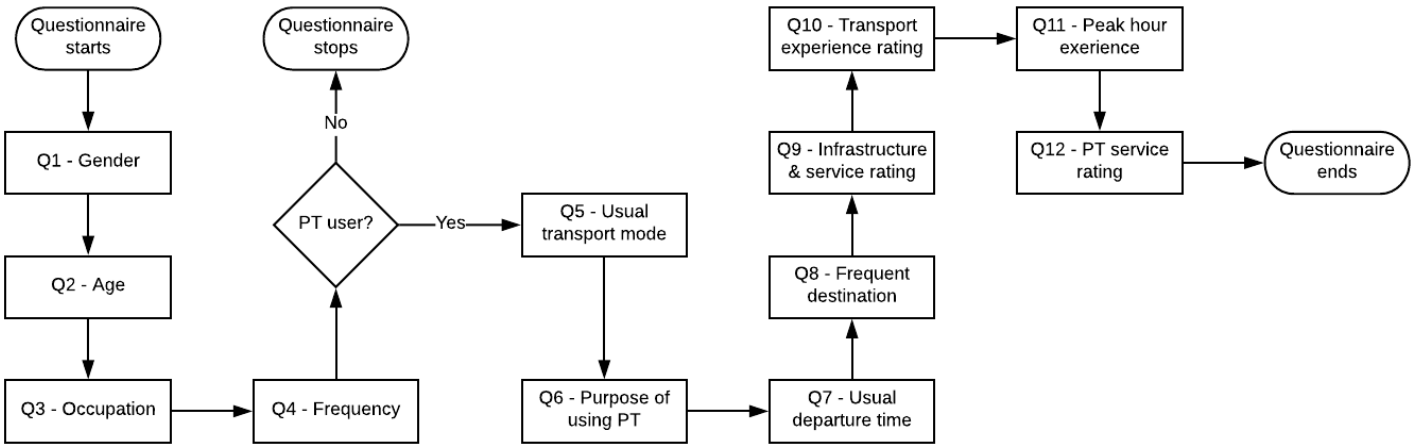

Figure 1 - Questionnaire flow

As can be seen in the diagram, general questions are asked in the first 3 questions, which are followed with the fourth question on the frequency of PT use. This question declares the decision where the respondents who do not use PT service will stop answering the questionnaire; however, the responses are still recorded for further analysis. After proceeding with how frequent they use PT for commuting as the conclusion for the first part of the questionnaire, the respondents will be asked 8 more questions, which are arranged in an order resembling to the process of their experience in using PT and divided into 2 other parts. Firstly, the respondents will have 4 choices representing the NSW PT modes to evaluate their frequent choice of PT. After that, the next question identifies their reason or purpose when using PT since it is very important to classify purpose of mode choice. Coming up next is the question on when the respondents usually depart to their destination, followed by the next question on their usual destination when using PT service. After the first 4 question of the second part, the third part of the questionnaire aims to collect the opinions as well as ratings of the respondents about their experience when using PT service, which is organised into 3 groups, the infrastructure and service experience, the travel experience, and peak

hour travel experience. The last question of the questionnaire focuses on figuring out how the respondents feel about the overall PT service they are offered so that all responses will be classified later using the result of this question.

Last but not least, the validity of the responses is determined by the completeness of the responses as well as the relevance of the responses to the questions. There are a few cases that must be considered during the filtering process of the completed questionnaires as this is the most important step which will decide the quality of the qualitative result.

- The first case involves in the respondents who answer that they never use PT service but still proceed with later questions; as a result, the answers after question 4 are deleted as they may be accidental answers, but the first 4 questions are kept as they are not redundant.
- The second case involves in the respondents who answers are unreasonable. An example for this case is that a respondent answers that he is within 25-34 age group but a retiree since it is impossible to be a retiree at a very young age; subsequently, these questionnaires are discarded as they represent the redundancy.
- The third case involves in the respondents who provide multiple answers in questions requiring a single answer; accordingly, these questionnaires are disregarded as there is no way to validate the answer.

Therefore, the whole dataset acquired from the web-based survey will be programmatically cleaned using the validation concept above to clean the data before putting into analysis. As a remark, RStudio is ideal for data cleaning and visualisation, free and open-sourced, as well as more lightweight than other data analysis tools; so, this integrated development environment (IDE) for R, a programming language for statistical computing and graphics, will be used in this qualitative design to handle the data from the survey. From the survey data, some preliminary hypotheses will be examined in order to associate with the later part of quantitative research to draw a solid conclusion.

III. RESULTS & DISCUSSION

The survey was conducted during the period of 2 weeks from 15 July 2018 to 31 July 2018 with 500 invitations sent to friends and acquaintances of the project group members. Out of 500 invitations, there were only 168 results from the invitees. The summary of the survey result is illustrated in the following Table 3:

Total invitations	500
Returned surveys	144
% of completion	33.60%
Valid surveys	121
Minimum surveys required	329
% of achievement	36.78%

Table 3 - Survey result summary

The failure of the survey contributes to the fact that the survey was conducted in a short amount of time (2 weeks) and the willingness of the invitees is not so high since the survey period occurred during the time that they were busy preparing for the new fiscal year, or the new semester as per reply from some respondents. However, the survey analysis is still conducted so that some useful information can be extracted from it.

At the outset, the proportions of respondents in terms of age, frequency of PT use, gender, occupation, as well as transport mode use are investigated in order to understand the distribution of respondent types, which is shown the Figure 2 below.

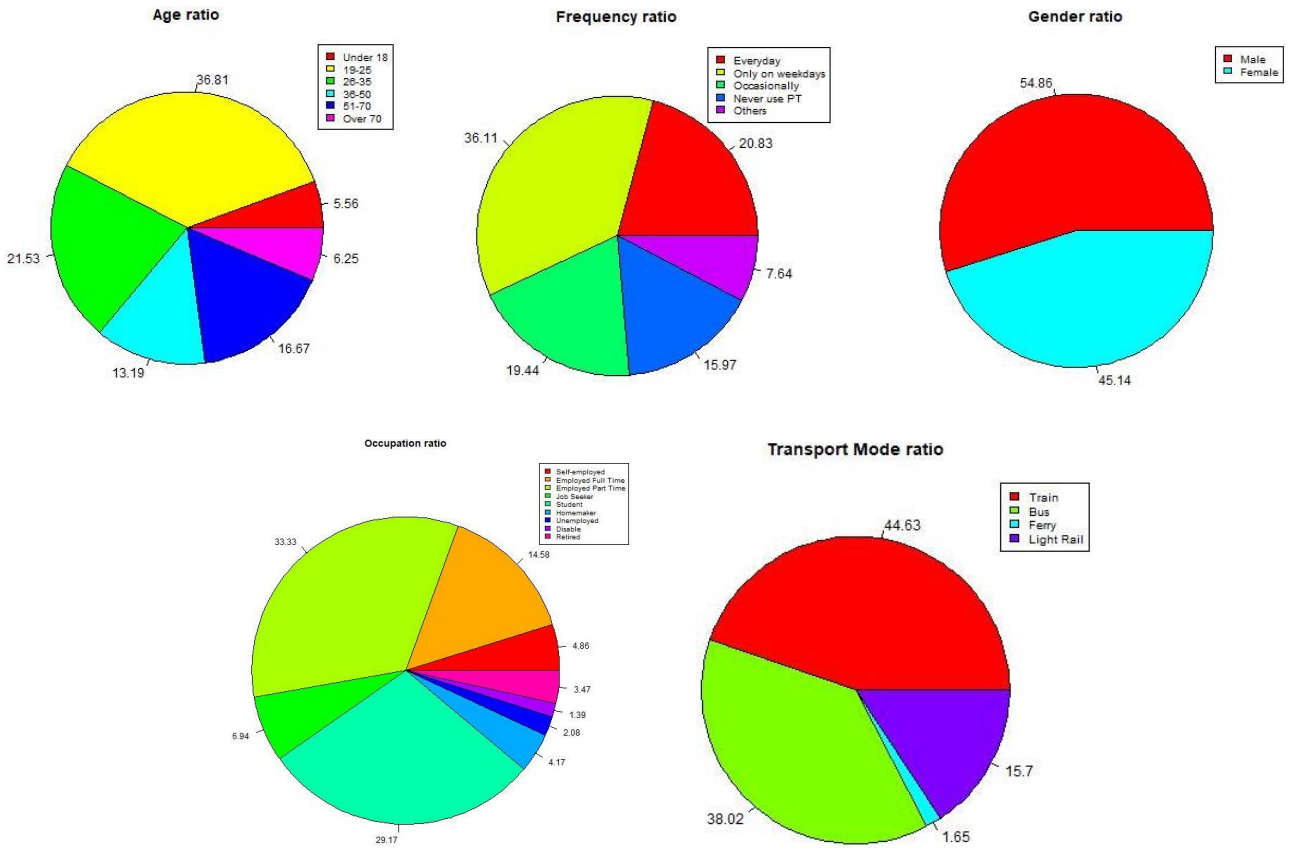

Figure 2 - Question 1 to 5 summary

A more concise summary can be illustrated in the following Table 4:

No.	Question statement	Offered response	Percentage (N = 121)
1	What is your gender?	Male	54.86%
		Female	45.14%
2	What age group do you belong to?	Under 18	5.56%
		19-25	36.81%
		26-35	21.53%
		36-50	13.19%
		51-70	16.67%
		Over 70	6.25%
3	What is your occupation?	Self-employed	4.86%
		Employed Full Time	14.58%
		Employed Part Time	33.33%
		Job Seeker	6.94%
		Student	29.17%
		Homemaker	4.17%
		Unemployed	2.08%
		Disabled	1.39%
		Retired	3.47%

4	How often do you use PT? *(The response "I never user PT" was made before the number of valid results was 144.)*	Everyday	20.83%
		Only on weekday	36.11%
		Occasionally	19.44%
		I never use PT	15.97%
		Others	7.64%
5	What is your usual PT mode when using PT?	Train	44.63%
		Bus	38.02%
		Ferry	1.65%
		Light rail	15.7%

Table 4 - Survey summary from question 1 to 5

The summary in Table 4 shows that the gender ratio of the survey is somewhat balanced with a slightly more male than female, over 70% of respondents are within the age of labour force (19 to 50 years old), about 10% of respondents are people that are not available for working, most of the respondents (about 85%) use PT service, and the most popular PT modes are train and bus.

In addition, the co-relations between each pair of questions from 1 to 5 are also illustrated by the following figures:

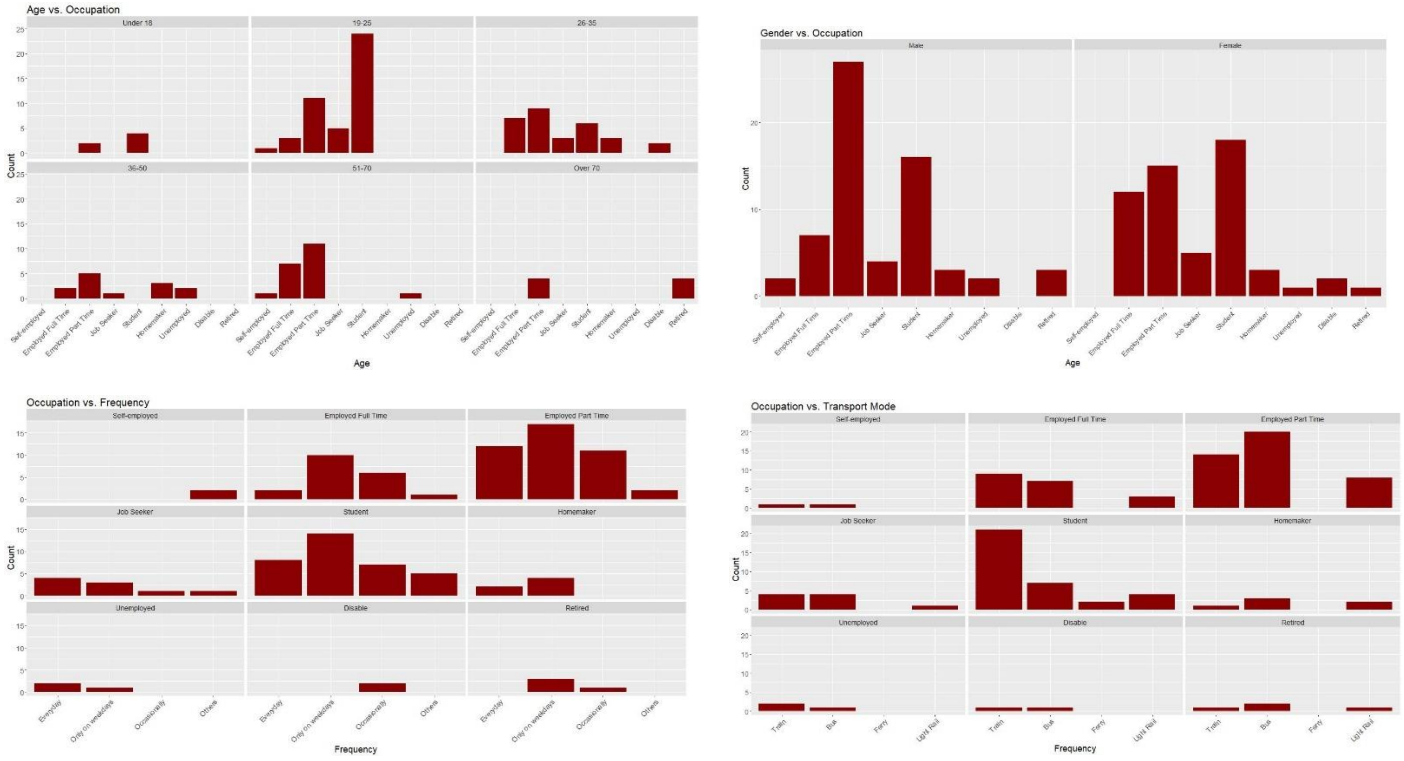

Figure 3 - Co-relations between variable pairs

It can be seen from Figure 3 that respondents ages from 19-25 are mostly part-timers and students, and they are mostly male. Additionally, they are also the respondents with somewhat frequency in PT use aside from full-timers, and students and full-timers are people who usually use train while part-timers usually choose both bus and train.

Moreover, the result of responses of Question 6 shows that PT users usually use PT service for:

- getting to work,
- go to school/university/educational institute,

- going shopping,
- family outing, and
- travelling to the CBD.

From there, it can be seen that the purpose of choosing to travel by PT varies across respondents, but the most common pattern for their choice is that it serves their normal daily life activities. As such, it is believed that PT has been becoming more and more popular among citizens as a means to replace their private vehicles. However, they usually do not feel comfortable when using PT during peak hour, which spans from 7 AM to 9 AM in the morning and 4 PM to 6:30 PM in the afternoon, as the respondents describe their experience in Question 11 when travelling during peak hour as terrible, in which they:

- have to stand most of the time even when travelling long distance,
- have to stand in crowded carriages without much space to breath,
- feel nervous just at the start of the day due to over-crowded carriages, and
- feel it difficult to keep their cool when standing next to some annoying passengers who do not care about their surroundings

However, when being asked about the usual departure time, most respondents reply that they normally travel during peak hour, particularly after 7:30 AM, while a portion of them reply that their departure time is not fixed, it changes based on their demand. As well, the respondents also signify that they mostly use PT service when needed, not on a regular basis, however, this principal does not apply for respondents who are student, part-timers, and full-timers, who choose PT for specific purpose. Moreover, most respondents reply that their usual destination are Redfern, Town Hall, Wolli Creek, and North Wollongong stations. This response is rather expected given to the fact that most respondents are UoW students, while some of them are employees, only a small portion of respondents are random.

Upon checking the responses for Question 9 and 10, which is illustrated as Figure 4 below, it can be seen that most respondents with different occupations react differently when being asked about their evaluation on the infrastructure and service quality. However, the responses from full-timers, part-timers, job seekers and students are somewhat stable, and it shows that they well perceive the

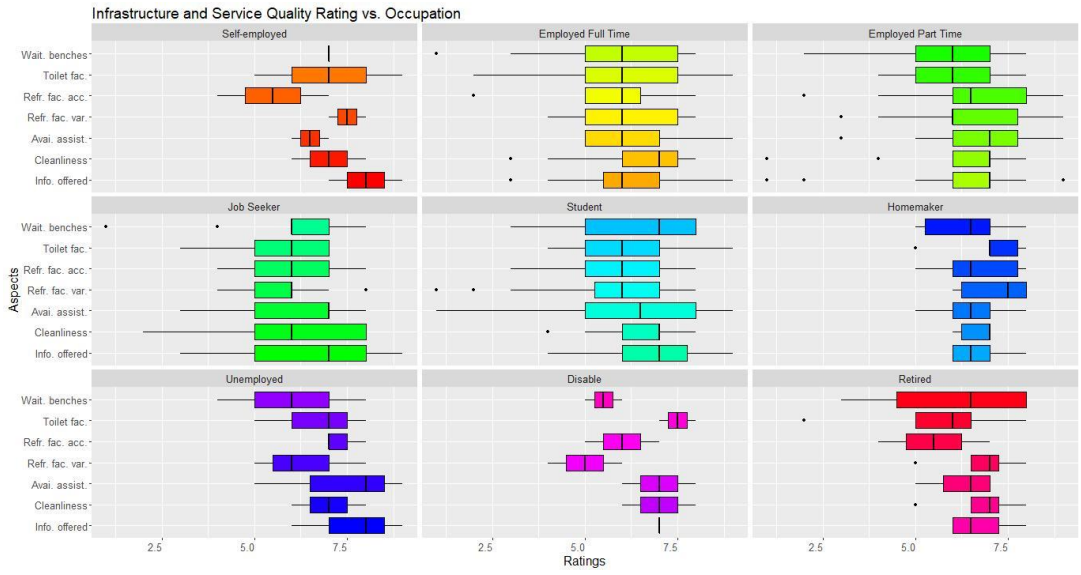

Figure 4 - Infrastructure and Service Quality evaluation in terms of occupation

quality of the service and infrastructure that are offered to them since they are the PT users who use PT service the most. Meanwhile, the value perception of other respondents with other occupations fluctuate a lot, which shows that there are some aspects that they do not feel satisfied with and some aspects give them the most satisfactory feeling.

As well, Figure 5 shows the evaluation of transport experience when separating with occupations. It can be seen that the respondents are rather satisfied with the temperature and the cleanliness of the transport carriage(s) as well as the comfort during their travel, while most of them do not feel really satisfied with other passengers' behaviours, or on-board information that is offered to them.

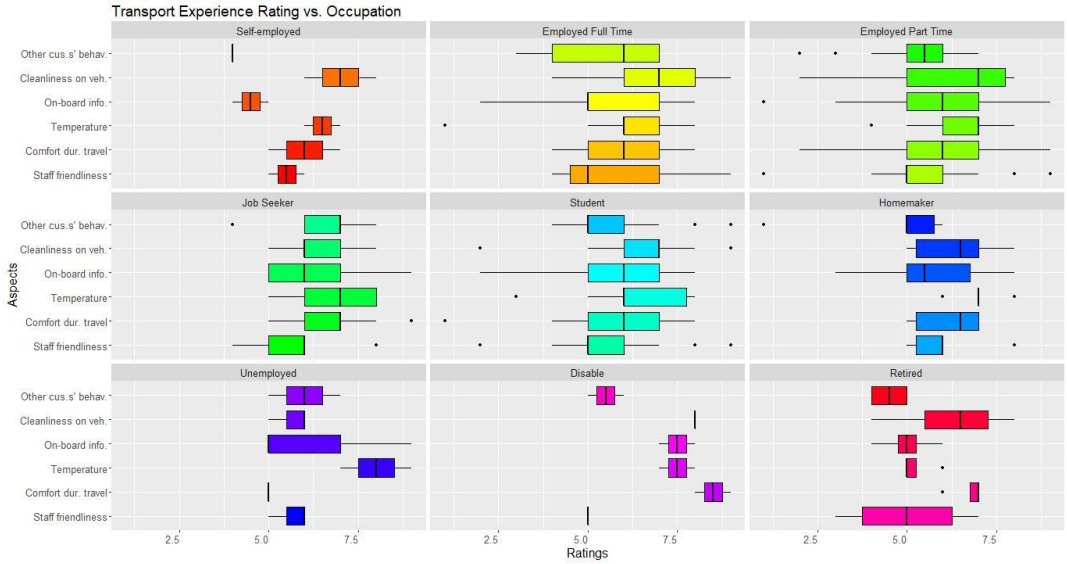

Figure 5 - Transport Experience evaluation in terms of occupation

Lastly, when being asked to give evaluation about overall experience when using PT service, most of the respondents evaluate it as somewhat above average but definitely not excellent enough, which is illustrated in Figure 6 below.

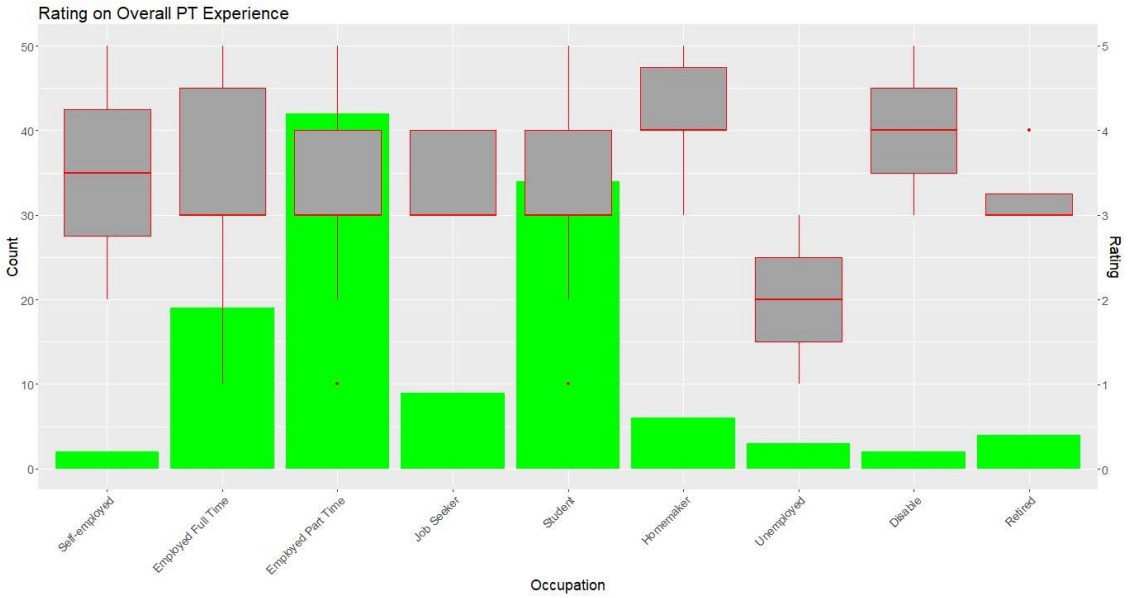

Figure 6 - Overall PT experience evaluation

It can be seen that most respondents feel that the current PT service is improving, but not to the point that they may evaluate highly due to the fact that their transport experience, which has been mentioned, is not good experience.

The result of this research suggests that the values that PT users perceive base on the frequency of their choice. As well, the purpose that PT users choose to shift from private vehicle to PT may be varied, but the most conspicuous one contributes to the fact that the PT network has been expanded vigorously in the last 5 years, as well as the pricing for Opal trips has become more reasonable comparing to the fuel price given to the fact that fuel price has become unstable in the last 5 years due to international conflicts and insecure political situations. Moreover, although the respondents feel uncomfortable when travelling in PT during peak hours, when being asked, they still reply that they still chose to travel in PT since they know the specific time that they will arrive at the destination, which is very convenient for them to plan for their activities. Out of 7 available responses offered in Question 9, the respondents replied positively on all of them, except for the available assistance and the cleanliness. This may result from their experience when travelling through some busy stations, which is very big in size but with scarce labour force. Therefore, investigating into the deployment of staff during each time frame or the possibility of employing new staff is necessary to improve the quality of PT service. However, since the qualitative research did not reach the target of sample size, the result is somewhat unstable and must be re-conducted when the time and human resource allow. As such, the quantitative research presented after this qualitative research will be able to uncover some uncertainty within the result of qualitative research since the data used in this research has been collected through a long time and will represent some pattern that cannot be figured out using qualitative data.

B. QUANTITATIVE DESIGN AND DATA ANALYSIS

I. INTRODUCTION

The quantitative part of this research focuses on analysing the data given by the research supervisor, which captures the information of the PT users every time they use their cards to tap on/off the terminal to proceed to or exit from the PT stations. The time that the records were captured is during August and September 2016. In this design, the data will be processed first in order to clean some unexpected or unnecessary data points before proceeding to analyse it; then, some data processing and analytic tools (i.e. Hadoop, Tableau...) will be utilised to showcase the underlying meaning of the data. The purpose of this part is to analyse the real-time data collected from the activities of the PT users so that their behaviours will be discovered through the visualisation of the data analysis. The result of this part will be coupled with the result of qualitative part for an overall discussion on the travel behaviour of PT users.

II. METHODOLOGY

The given dataset consists of two sets of datasets collected from Opal card users in 50 days from 01 August 2016 to 19 August 2016, one is weekly data about the Opal card registered with TfNSW including the single trip tickets, the other is data about the trips they made during this period. The latest card dataset consists of 12,957,955 records about Opal card users, while the total records the journey dataset offers clock the amount of 94,521,851 (which implies that an average of about 1,890,437 records were collected daily). Upon examining the card dataset, there are some problems with the card dataset as it consists of some fields with abnormal values such as “UNKNOWN”, which can be considered as data collection error, and “BLOCKED”, “EXPIRED”, “HOTLISTED”, “DEHOTLISTED”, or “INITIALISED”, which can be considered as manipulated data. As well, there are lots of abnormal values in journey dataset, which is designated as “UNKNOWN” or “-1”. Therefore,

data pre-processing is required before proceeding with data analysis. However, in order to handling such large amount of data, which is combined into 85Gb in size, a big data processing tool must be employed; thus, Apache Spark which stems from Hadoop ecosystem has been used to handle the task since it is capable of turning and processing data in batch which does not rely heavily on hardware resources and its query programming syntax is similar to SQL query which is familiar to the project team members.

Firstly, the card dataset will be dealt with before proceeding to the journey dataset since it can be done more quickly, its schema is as follows:

```
root
 |-- CARD_SK: integer (nullable = true)
 |-- CIN: string (nullable = true)
 |-- CARD_STAT_CD: string (nullable = true)
 |-- REGD_CARD_IND: string (nullable = true)
 |-- OWNR_PSTCD: integer (nullable = true)
 |-- OWNR_BRTH_YR: integer (nullable = true)
 |-- CARD_TYP_SK: integer (nullable = true)
 |-- CARD_TYP_CD: string (nullable = true)
 |-- SMART_CARD_TYP_CD: string (nullable = true)
 |-- PSNGR_TYP_SK: integer (nullable = true)
 |-- PSNGR_TYP_DESC: string (nullable = true)
 |-- DISC_ENT_SK: integer (nullable = true)
 |-- DISC_ENT_CD: integer (nullable = true)
 |-- DISC_ENT_DESC: string (nullable = true)
```

The distinct values of the discount entitlement, card status, as well as card type fields (the single trip tickets are combined together for cleaner explanation) are as follows:

```
+-------------+
|DISC_ENT_DESC|
+-------------+
|        No DE|
|  No Discount|
|   Concession|
|    Pensioner|
+-------------+
```

```
+------------+
|CARD_STAT_CD|
+------------+
|     EXPIRED|
| DEHOTLISTED|
|   HOTLISTED|
|     BLOCKED|
| INITIALISED|
|     ENABLED|
|     UNKNOWN|
+------------+
```

```
+----------------+
|CARD_TYP_CD     |
+----------------+
|Senior/Pensioner|
|Adult           |
|UNKNOWN         |
|Free Travel     |
|School Student  |
|Child/Youth     |
|Employee        |
|Sgl Trip Ticket |
|Concession      |
+----------------+
```

The distinct association of values between the card type and discount entitlement fields, card type and passenger type fields, as well as card status and discount entitlement fields are as follows:

```
+----------------+--------------+
|CARD_TYP_CD     |PSNGR_TYP_DESC|
+----------------+--------------+
|Adult           |Child/Youth   |
|Senior/Pensioner|Adult         |
|Sgl Trip Ticket |Adult         |
|Sgl Trip Ticket |Child/Youth   |
|Concession      |Adult         |
|Employee        |Adult         |
|Free Travel     |Adult         |
|UNKNOWN         |Adult         |
|School Student  |Adult         |
|Child/Youth     |Child/Youth   |
|Adult           |Adult         |
|Child/Youth     |Adult         |
+----------------+--------------+
```

```
+----------------+-------------+
|CARD_TYP_CD     |DISC_ENT_DESC|
+----------------+-------------+
|Employee        |Concession   |
|Sgl Trip Ticket |No Discount  |
|Senior/Pensioner|Pensioner    |
|Adult           |No Discount  |
|Adult           |Concession   |
|School Student  |No Discount  |
|UNKNOWN         |No Discount  |
|Senior/Pensioner|No Discount  |
|Concession      |No Discount  |
|Adult           |Pensioner    |
|Adult           |No DE        |
|Concession      |Concession   |
|Free Travel     |Concession   |
|School Student  |Concession   |
|Child/Youth     |No Discount  |
+----------------+-------------+
```

```
+------------+-------------+
|CARD_STAT_CD|DISC_ENT_DESC|
+------------+-------------+
|INITIALISED |No Discount  |
|DEHOTLISTED |Pensioner    |
|ENABLED     |No DE        |
|BLOCKED     |No Discount  |
|ENABLED     |Pensioner    |
|BLOCKED     |Concession   |
|HOTLISTED   |No Discount  |
|BLOCKED     |Pensioner    |
|EXPIRED     |No Discount  |
|ENABLED     |No Discount  |
|HOTLISTED   |No DE        |
|UNKNOWN     |Pensioner    |
|DEHOTLISTED |No DE        |
|HOTLISTED   |Pensioner    |
|UNKNOWN     |No Discount  |
|DEHOTLISTED |No Discount  |
|DEHOTLISTED |Concession   |
|ENABLED     |Concession   |
|BLOCKED     |No DE        |
|HOTLISTED   |Concession   |
|UNKNOWN     |Concession   |
+------------+-------------+
```

As can be seen, the problems in the card dataset lie within the abnormal values; however, upon checking the journey dataset, it appears that some cards which statuses are marked as “UNKNOWN”, “BLOCKED”, “HOTLISTED”, “EXPIRED” ... or types are marked as “UNKNOWN” can still be used to travel in PT network. Thus, it can be safely assumed that all of them are data collection error, which can be fixed and used normally without deletion. There is also a problem with the records inside the card dataset, which is that the card type is displayed as “Child/Youth” while the passenger type is displayed as “Adult” and the discount entitlement consists of two same values, “No Discount” and “No DE” (discount entitled); as such, these abnormalities are also considered as data collection error and need fixing. In order to fix this problem, the following assumptions are employed:

1. Firstly, if the discount entitlement (DISC_ENT_DESC) is “No DE”, it will be changed to “No Discount”.
2. After that, all “Single Trip ...” types (CARD_TYP_CD) will be converted to “Single Trip Ticket” using regular expression “Sgl Trip ([A-Za-z]+) ([A-Za-z]+).*” as they represent PT users who don’t use Opal cards.
3. Then, if the card type (CARD_TYP_CD) is “UNKNOWN” and the discount entitlement (DISC_ENT_DESC) is “No Discount”, that card will be assumed as “Adult” type; other than that, the card types match the discount entitlements and other discount entitlements don’t associate with “UNKNOWN” card type, so no change is needed.
4. Nest, if the card statuses (CARD_STAT_CD) are of abnormal values (“UNKNOWN”, “HOTLISTED”, “BLOCKED”, “EXPIRED” ...), they will be changed to “TEMPORARY” status as it is not certain whether the cards are used permanently or not, or whether they are discarded or not.
5. Then, if the card type (CARD_TYPCD_ is “Child/Youth” and the passenger type (PSNGR_TYP_DESC) is “Adult” as well as the discount entitlement (DISC_ENT_DESC) is “No Discount”, that passenger will be assumed as “Child/Youth” type instead of “Adult” since “No Discount” entitlement is associated with both “Adult” and “Child/Youth” card types but “Adult” card type also associates with many other discount entitlement while “Child/Youth” card type only associates with “No Discount” so it may be safer to assume the passenger to be “Child/Youth” rather than “Adult” type.
6. Lastly, if the card type (CARD_TYP_CD) is “UNKNOWN” and the card status (CARD_STAT_CD) is also “UNKNOWN”, they will be discarded since there is no certain way to ascertain the assumption made for those card records. The reason for this action is because query upon this situation returns all information with unknown values as shown in the below result, of which no hints can be figured out based on the card owners’ birth years or postcodes in order to make an assumption; furthermore, the total amount of these cards are 23,698, which makes up about 0.18% total amount of records, so it won’t cause much impact to the dataset after being deleted.

```
+------------+-----------+-------------+----------+------------+-----+
|CARD_STAT_CD|CARD_TYP_CD|DISC_ENT_DESC|OWNR_PSTCD|OWNR_BRTH_YR|count|
+------------+-----------+-------------+----------+------------+-----+
|     UNKNOWN|    UNKNOWN|  No Discount|        -1|          -1|23698|
+------------+-----------+-------------+----------+------------+-----+
```

On the other hand, there exist so many problems within the journey dataset due to data collection error, which can be signified as four major problems; so, in order to easily identify the abnormalities of the data records, some unnecessary fields will be discarded, the following fields are retained:

```
root
 |-- CARD_FK: integer (nullable = true)
 |-- PSNGR_TYP_CD: string (nullable = true)
 |-- CARD_TYP_CD: string (nullable = true)
 |-- JRNY_ID: integer (nullable = true)
 |-- DISC_ENT_DESC: string (nullable = true)
 |-- TS_TYP_CD: string (nullable = true)
 |-- JS_DURN_SEC: integer (nullable = true)
 |-- IMTT_DESC: string (nullable = true)
 |-- TAG1_DT_FK: integer (nullable = true)
 |-- TAG2_DT_FK: integer (nullable = true)
 |-- TAG1_TM: string (nullable = true)
 |-- TAG2_TM: string (nullable = true)
 |-- TAG1_TS_NM: string (nullable = true)
 |-- TAG1_LAT_VAL: double (nullable = true)
 |-- TAG1_LONG_VAL: double (nullable = true)
 |-- TAG2_TS_NM: string (nullable = true)
 |-- TAG2_LAT_VAL: double (nullable = true)
 |-- TAG2_LONG_VAL: double (nullable = true)
 |-- TAG1_OFF_PEAK_IND: boolean (nullable = true)
 |-- TAG2_OFF_PEAK_IND: boolean (nullable = true)
```

The first problem of this dataset comes from a mixture of a record in July, which was on 18 July 2016 and 31 July 2016 as show in the below two query results. It can be seen in the first query result that there is only one record that was recorded to tap-on on 18 July 2016 but tap-off on 30 August 2016 with the trip duration was counted as -2,760 seconds (which only applies for trips that are not tapped-off at the end of the day or without tap-on details in some stations) and the destination and departure are the same; thus, it will be discarded since there is no telling this record is accounted as valid or not and it is the sole record with this kind of value, which will not affect the analysis even if it disappears. Besides, there are some more problems coming from the records which tap-on date was on 31 July 2016 and tap-off date was on 01 August 2016. As can be seen in the second query result, these records have tap-on date and time, tap-off date and time as well as tap-on and -off locations, so they can be easily checked to see whether the time is valid for the trip duration or not. Take the first row of the second data frame as an example, the distance between the Woy Woy Station and Town Hall Station is about 75 km and it takes about one and a half hour at most to travel between them by train, another example that can be verified lies within the 5th row of the second query result which takes about 1 hour 50 minutes at most to travel from Wollongong Station to Domestic (Airport) Station by train; so that makes sense to the tap-on and -off time shown in the data frame. Therefore, it is assumed that this is data collection error and can be fixed by changing the tap-on date (TAG1_DT_FK) to be the same as tap-off date (TAG2_DT_FK), then subtract the tap-on and -off time to reconstruct the duration of journey within these 2 fields. The same principal can be applied for trips that have the same problem, so a general assumption can be safely made using this strategy.

PSNGR_TYP_CD	TS_TYP_CD	TAG1_DT_FK	TAG2_DT_FK	JS_DURN_SEC	TAG1_TS_NM	TAG2_TS_NM
Adult	Train	20160718	20160830	-2760	Rockdale Station	Rockdale Station

```
+------------+---------+----------+--------+----------+--------+-----------+--------------------+--------------------+
|PSNGR_TYP_CD|TS_TYP_CD|TAG1_DT_FK| TAG1_TM|TAG2_DT_FK| TAG2_TM|JS_DURN_SEC|          TAG1_TS_NM|          TAG2_TS_NM|
+------------+---------+----------+--------+----------+--------+-----------+--------------------+--------------------+
|       Adult|    Train|  20160731|04:30:00|  20160801|06:03:00|     -80820|     Woy Woy Station|   Town Hall Station|
|       Adult|    Train|  20160731|04:14:00|  20160801|05:46:00|     -80880|    Tuggerah Station|Macquarie Univers...|
|       Adult|    Train|  20160731|04:29:00|  20160801|05:03:00|     -84360|   Riverwood Station|    Domestic Station|
|       Adult|    Train|  20160731|04:21:00|  20160801|05:24:00|     -82620| Seven Hills Station| St Leonards Station|
|       Adult|    Train|  20160731|04:26:00|  20160801|06:00:00|     -80760|  Wollongong Station|    Domestic Station|
|       Adult|    Train|  20160731|04:27:00|  20160801|05:04:00|     -84180|    Lidcombe Station|   Town Hall Station|
|       Adult|    Train|  20160731|04:28:00|  20160801|04:37:00|     -85860|Warwick Farm Station|  Cabramatta Station|
|       Adult|    Train|  20160731|04:15:00|  20160801|04:45:00|     -84600|  Flemington Station|   Punchbowl Station|
|       Adult|    Train|  20160731|04:30:00|  20160801|06:48:00|     -78120|   Fassifern Station| Strathfield Station|
|       Adult|    Train|  20160731|04:22:00|  20160801|05:27:00|     -82500|  Merrylands Station|   Macarthur Station|
|       Adult|    Train|  20160731|03:58:00|  20160801|05:21:00|     -81420|Pennant Hills Sta...|    Domestic Station|
|       Adult|    Train|  20160731|04:28:00|  20160801|05:56:00|     -81120|     Panania Station|   Chatswood Station|
|       Adult|    Train|  20160731|03:38:00|  20160801|06:59:00|     -74340| Broadmeadow Station|     Central Station|
|       Adult|    Train|  20160731|04:29:00|  20160801|05:49:00|     -81600| Canley Vale Station|       Tempe Station|
|       Adult|    Train|  20160731|04:20:00|  20160801|05:12:00|     -83280|      Casula Station|  Parramatta Station|
|       Adult|    Train|  20160731|04:28:00|  20160801|05:08:00|     -84000|     Yagoona Station|    Sydenham Station|
|       Adult|    Train|  20160731|04:25:00|  20160801|05:18:00|     -83220|  Parramatta Station|Circular Quay Sta...|
|       Adult|    Train|  20160731|04:21:00|  20160801|05:25:00|     -82560|   Blacktown Station|     Banksia Station|
|       Adult|    Train|  20160731|04:29:00|  20160801|06:01:00|     -80880|     Leumeah Station|North Sydney Station|
|       Adult|    Train|  20160731|04:12:00|  20160801|06:16:00|     -78960|       Wyong Station|North Sydney Station|
+------------+---------+----------+--------+----------+--------+-----------+--------------------+--------------------+
only showing top 20 rows
```

The second problem of the journey dataset is somewhat similar to the card dataset, since the Opal card number is a foreign key in the journey dataset while it is the primary key in the card dataset, where there exist some abnormal values in the card type field as well as discount entitlement field just like in the card dataset; moreover, the “Single Trip ...” types also appear in the journey dataset as the data was collected indiscriminately, so it has the same issue in the journey dataset as in the card dataset. Thus, the same strategy for data processing in the card dataset will be inherited to the journey dataset.

The third problem of the journey dataset is related to the inter-modal transfer indicator since there are some records missing this indicator (which are marked as “UNKNOWN”) and it will affect the data analysis. As such, if the journey record is a valid record, which has valid tap-on and -off location as well as date and time, the inter-modal transfer indicator can be re-evaluated using the 60 minutes transfer duration (any PT transport mode change occurring within 60 minutes after finishing the previous trip can be counted as a transfer) if the transit types between 2 consecutive trips are known and different.

The fourth problem of the journey dataset lies within the unknown tap-on and -off date and time of the journey records (which is marked as “-1”, a.k.a. “UNKNOWN”) as per below two query results. Per unknown tap-on date and time, since it leads to tap-on location to be marked as “UNKNOWN” and the journey duration is counted as negative as shown in the first query result, there is no assured means to ascertain the assumption to fix these records, it may be safe to remove them if there is good enough reason. As for unknown tap-off date and time, since all of these records lead to journey duration to be counted as 0 and tap-off location to be shown as “UNKNOWN” as per second query result, it can be assumed that the tap-on and -off location as well as date and time are the same, which is very normal when a passenger taps on but decides to change mind later and taps on reverse to nullify the previous tap-on in order not to incur all-day charge without using the service.

```
+------------+---------+----------+-------+----------+----------+--------+-----------------------------------+-----------+
|PSNGR_TYP_CD|TS_TYP_CD|TAG1_DT_FK|TAG1_TM|TAG1_TS_NM|TAG2_DT_FK|TAG2_TM |TAG2_TS_NM                         |JS_DURN_SEC|
+------------+---------+----------+-------+----------+----------+--------+-----------------------------------+-----------+
|Adult       |Bus      |-1        |-1     |UNKNOWN   |20160801  |04:47:00|Broadway before Mountain St        |-85380     |
|Adult       |Bus      |-1        |-1     |UNKNOWN   |20160801  |04:45:00|Lidcombe Station, Church St, Stand B|-84360     |
|Adult       |Bus      |-1        |-1     |UNKNOWN   |20160801  |04:34:00|Burwood Station, Railway Pde, Stand C|-84600     |
|Adult       |Bus      |-1        |-1     |UNKNOWN   |20160801  |04:47:00|Museum Station, Elizabeth St, Stand D|-85260     |
|Adult       |Bus      |-1        |-1     |UNKNOWN   |20160801  |04:42:00|Oxford St near Verona St           |-85440     |
|Adult       |Bus      |-1        |-1     |UNKNOWN   |20160801  |04:36:00|Bondi Junction Interchange (Set Down)|-85920     |
|Adult       |Bus      |-1        |-1     |UNKNOWN   |20160801  |04:47:00|Museum Station, Elizabeth St, Stand D|-85200     |
|Adult       |Bus      |-1        |-1     |UNKNOWN   |20160801  |04:52:00|Circular Quay, Young St, Stand D   |-85080     |
|Adult       |Bus      |-1        |-1     |UNKNOWN   |20160801  |04:35:00|Oxford St near Bronte Rd           |-85800     |
|Adult       |Bus      |-1        |-1     |UNKNOWN   |20160801  |04:45:00|Oxford St near Brisbane St         |-85380     |
+------------+---------+----------+-------+----------+----------+--------+-----------------------------------+-----------+
only showing top 10 rows

+------------+---------+----------+--------+-----------------------------------+----------+-------+----------+-----------+
|PSNGR_TYP_CD|TS_TYP_CD|TAG1_DT_FK|TAG1_TM |TAG1_TS_NM                         |TAG2_DT_FK|TAG2_TM|TAG2_TS_NM|JS_DURN_SEC|
+------------+---------+----------+--------+-----------------------------------+----------+-------+----------+-----------+
|Adult       |Bus      |20160801  |04:53:00|Lakemba Station, The Boulevarde    |-1        |-1     |UNKNOWN   |0          |
|Adult       |Bus      |20160801  |08:54:00|Fairfield Interchange - Stand F    |-1        |-1     |UNKNOWN   |0          |
|Adult       |Bus      |20160801  |06:59:00|Junction Rd near Renton Av         |-1        |-1     |UNKNOWN   |0          |
|Adult       |Bus      |20160801  |06:00:00|Central Station, Eddy Ave, Stand C |-1        |-1     |UNKNOWN   |0          |
|Adult       |Bus      |20160801  |05:59:00|Central Station, Eddy Ave, Stand C |-1        |-1     |UNKNOWN   |0          |
|Adult       |Ferry    |20160801  |08:40:00|Stockton Wharf                     |-1        |-1     |UNKNOWN   |0          |
|Adult       |Ferry    |20160801  |10:51:00|Stockton Wharf                     |-1        |-1     |UNKNOWN   |0          |
|Adult       |Ferry    |20160801  |13:18:00|Stockton Wharf                     |-1        |-1     |UNKNOWN   |0          |
|Adult       |Ferry    |20160801  |09:04:00|Stockton Wharf                     |-1        |-1     |UNKNOWN   |0          |
|Adult       |Ferry    |20160801  |12:34:00|Stockton Wharf                     |-1        |-1     |UNKNOWN   |0          |
+------------+---------+----------+--------+-----------------------------------+----------+-------+----------+-----------+
only showing top 10 rows
```

Therefore, the following assumptions are made in order to reconstruct the journey dataset:

1. Firstly, if there exist journeys with transit type (TS_TYP_CD) as “UNKNOWN” as in the below query results or tap-on date (TAG1_DT_FK) as “-1” (or “UNKNOWN”), they will be removed since there are not enough hints to lead to a certain assumption to fix these records. The reason for this removal is that the unknown transit type (TS_TYP_CD) leads to unknown tap-on (TAG1_TS_NM) and -off (TAG2_TS_NM) locations so even though there may exist the tap-on (TAG1_DT_FK) or -off (TAG2_DT_FK) details, those journeys will not belong to any transport type and not be useful for analysis. As well, the unknown tap-on date (TAG1_DT_FK) leads to unknown tap-on location (TAG1_TS_NM) so even if the tap-off date (TAG2_DT_FK) exists, the trip cannot be verified as no departure location is identified. Moreover, these journeys are within 717,523 records (6,356 for unknown transit type and 711,767 for unknown tap-on date) of the journey dataset, which is about 0.76% of the total records, so removing them may not cause big impact to data analysis.

```
+------------+---------+----------+--------+----------+----------+--------+----------+-----------+
|PSNGR_TYP_CD|TS_TYP_CD|TAG1_DT_FK|TAG1_TM |TAG1_TS_NM|TAG2_DT_FK|TAG2_TM |TAG2_TS_NM|JS_DURN_SEC|
+------------+---------+----------+--------+----------+----------+--------+----------+-----------+
|Adult       |UNKNOWN  |20160801  |14:30:00|UNKNOWN   |20160801  |14:44:00|UNKNOWN   |840        |
|Adult       |UNKNOWN  |20160801  |14:29:00|UNKNOWN   |20160801  |14:44:00|UNKNOWN   |900        |
|Adult       |UNKNOWN  |20160801  |14:30:00|UNKNOWN   |20160801  |14:45:00|UNKNOWN   |900        |
|Adult       |UNKNOWN  |20160801  |14:30:00|UNKNOWN   |20160801  |14:44:00|UNKNOWN   |840        |
|Adult       |UNKNOWN  |20160801  |14:30:00|UNKNOWN   |20160801  |14:45:00|UNKNOWN   |900        |
|Adult       |UNKNOWN  |20160801  |14:30:00|UNKNOWN   |20160801  |14:45:00|UNKNOWN   |900        |
|Adult       |UNKNOWN  |20160801  |14:32:00|UNKNOWN   |20160801  |14:44:00|UNKNOWN   |720        |
|Adult       |UNKNOWN  |20160801  |14:30:00|UNKNOWN   |20160801  |14:44:00|UNKNOWN   |840        |
|Adult       |UNKNOWN  |20160801  |14:32:00|UNKNOWN   |20160801  |14:44:00|UNKNOWN   |720        |
|Adult       |UNKNOWN  |20160801  |14:30:00|UNKNOWN   |20160801  |14:44:00|UNKNOWN   |840        |
+------------+---------+----------+--------+----------+----------+--------+----------+-----------+
only showing top 10 rows
```

2. Then, the assumptions number 1, 2, 3, 5 of the card dataset will be reused in the journey dataset.
3. Next, when there exists unknown tap-off location as well as date and time values, since the journey duration is deemed as 0 by all means, the values of tap-on location together with date and time of the same row will be assumed for tap-off.
4. After that, when the tap-on date (TAG1_DT_FK) is “20160731”, it will be assumed to be the same as tap-off date (TAG2_DT_FK); else, when the tap-off date is different from the tap-on date and the total difference in time between the tap-on and -off time comparing to midnight

is more than 3 hours, it will be assumed to be the same of tap-on date since no single intra-state PT trip exceeds more than two and a half hours ; other than that, all tap-on and -off values are kept as-is.

5. Lastly, if there are journeys with inter-modal transfer indicator (IMTT_DESC) marked as "UNKNOWN", the journey will be ordered by the card number (CARD_FK), tap-on date (TAG1_DT_FK) and tap-on time (TAG1_TM) within the partition of tap-on date (TAG1_DT_FK). Then, the journeys with the duration more than 0 (exclusive) will be taken, the tap-on time of each row will be compared to the tap-off time of the preceding row when the card numbers of those two rows are the same; if the difference between those tap-on and -off times is under 60 (minutes) or 3,600 (seconds) and the 2 consecutive transit types are different, the unknown inter-modal transfer indicator will be reconstructed to appropriate inter-modal transfer type ("Rail to bus", "Bus to rail", "Bus to ferry", "Ferry to bus"...); otherwise, it will be assumed as "No inter-modal transfer".

After finishing reconstructing the journey duration of the journey dataset, the invalid journeys which duration is 0 (consists of 4,192,486 records) will be filtered again in order to complete the reconstructed journey dataset since they can be considered tap-on reversed records and do not make sense when visualise data on journeys that actually took place. As such, from the total of 94,521,851 records in the dataset, the total invalid records that are removed are 4,910,009 (5.19% of original dataset), which makes it the remaining total of 89,611,842 useful records (94.81% of original dataset) for data visualisation.

Finally, the information from the reconstructed journey and card datasets will be visualised by Tableau for data analysis after packing all query results into JSON files (for file size under 10Mb) or CSV files (for file size over 10Mb).

III. RESULTS & DISCUSSION

The result which is illustrated in Figure 7 shows that the weekly growth of Opal card users over the period of 9 weeks (from 07 August 2016 to 02 October 2016) is significant in the group of "Adult", "Child/Youth" and "Sgl Trip Ticket" while other card type shows less significant or almost no change, i.e. "Employee" and "School Student" card type. The most important point of this significance is that passengers tend to choose single trip ticket than applying for an Opal card, which is signified as the 274% growth after 9 observed weeks comparing to the 10.8% growth in adult passenger or 13.9% growth in child/youth passenger. Moreover, as of 02 October 2016, single trip tickets occupy 11% of the total collected data, which means that more than 10% of PT users prefer using paper ticket rather than Opal card. Meanwhile, the growth in the group of employee passengers shows slow increase (2.3% after 9 observed weeks), which implies that the Opal card is not so popular within salary people segment.

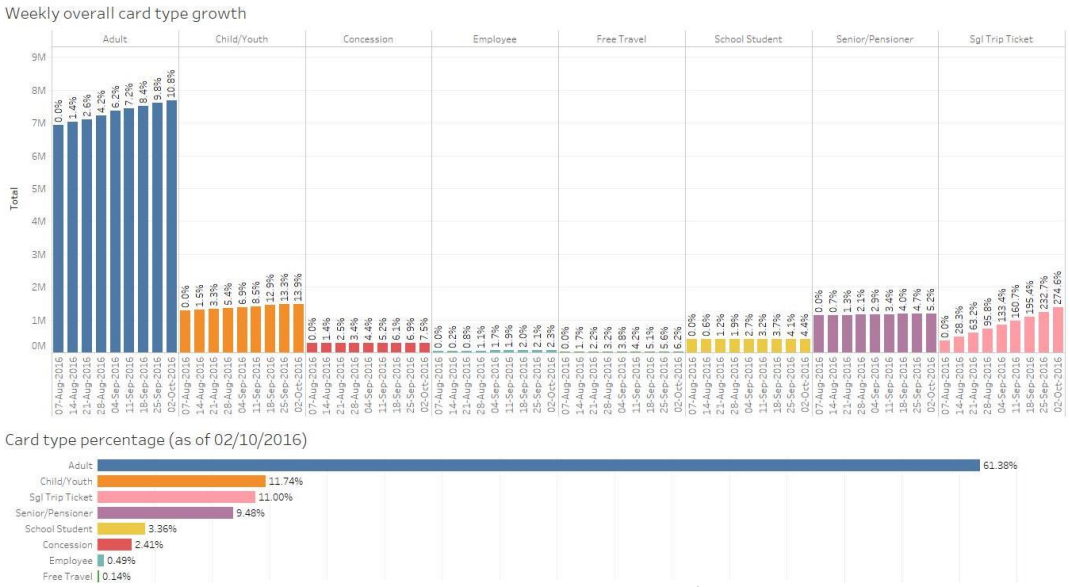

Figure 7 - Opal card growth in terms of card type

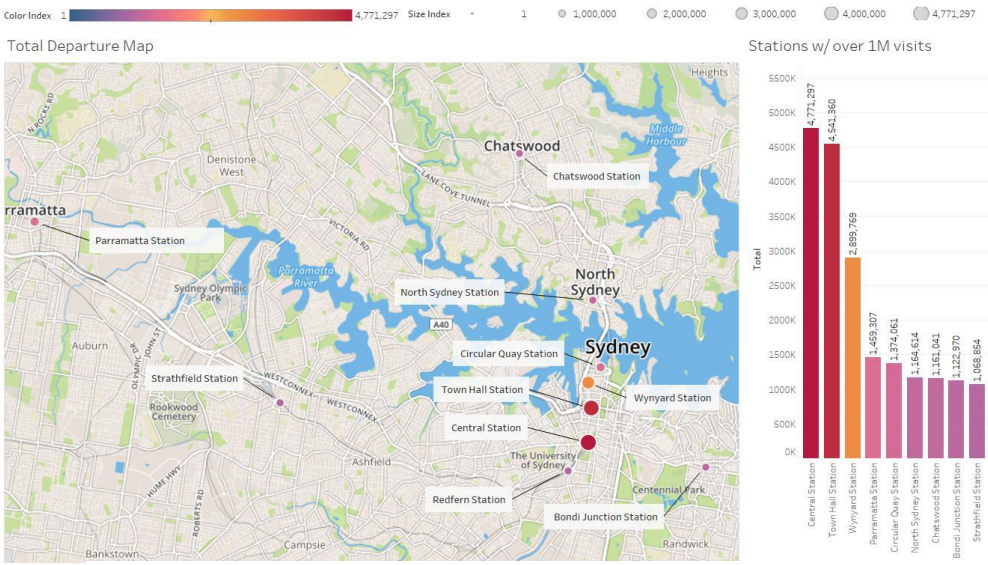

Figure 8 - Tap-on density from 01 August 2016 to 19 September 2016

On the other hand, the results illustrated in Figure 8 show 10 PT stations that served over one million passengers during the period of 01 August 2016 – 19 September 2016, which signifies that they served more than 20,000 passengers per day. As shown in the figure, the most outstanding PT stations belong to all train stations, which implies that train mode is the most popular PT mode choice. Furthermore, the detailed summary of PT activities during this period is illustrated as in Table 5 below.

	Off-peak period	Peak period	Total
Total trips made	20,481,573	69,130,269	89,611,842
Total journey duration	5,567,861 hrs	37,236,188 hrs	42,804,049 hrs
Avg time per trip	16 mins 19 secs	32 mins 19 secs	28 mins 40 secs
Total fare collected	AU$ 31,262,059.70	AU$ 160,055,350.20	AU$ 191,317,409.90
Avg fare paid per trip	AU$ 1.53	AU$ 2.32	AU$ 2.13

Table 5 - PT activity summary in period from 01 August 2016 to 19 August 2016

It can be seen that PT commuters pay for their trips during off-peak period less than 34% comparing to peak period; however, the transport volume as well as total transport duration during peak period

top off-peak period by about 3.37 times and 6.69 times respectively. Also, each PT commuter spends about 28 mins 40 seconds in average for each trip, but average travelling time during peak hour is almost two times longer than during off-peak hour. Thus, it is believed that the impact of off-peak and peak hour is very important to the established transport pattern of PT users. As such, the tap-on density during off-peak and peak hours will be examined to draw some hidden meaning behind the application of this concept, which can be seen from Figure 9, 10, 11, and 12.

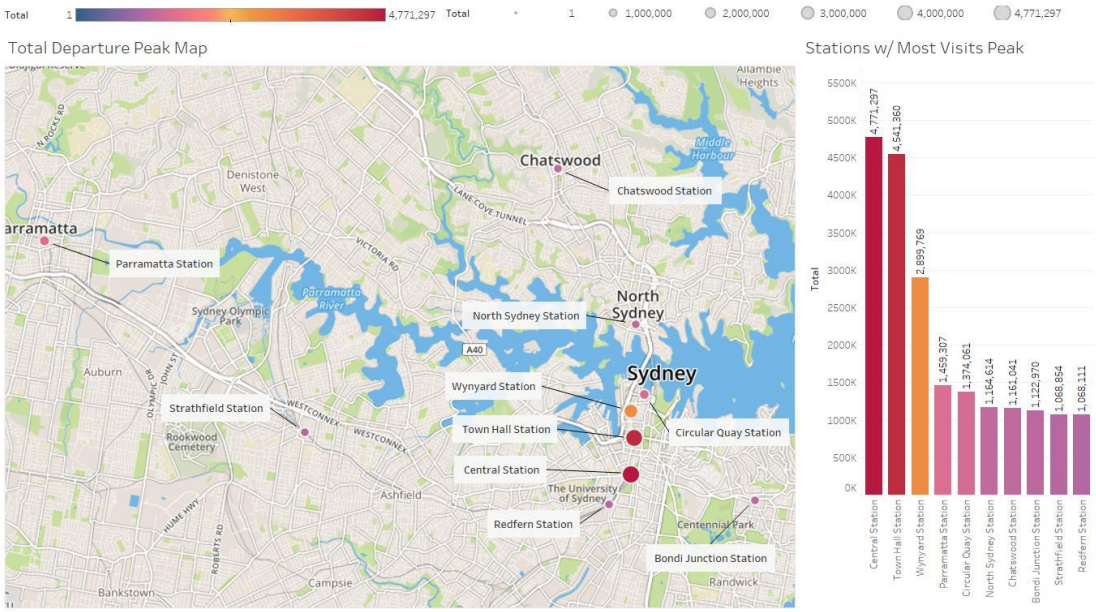

Figure 9 - Tap-on density during peak hour from 01 August 2016 to 19 September 2016

Per Figure 9, even though they are supposed to be two different figures, Figure 8 and 9 share the same visualisation, which means that train stations are the busiest stations during peak hour, and it also signifies that the PT lines that are connected to Sydney CBD (Redfern-Bondi Junction for T4 Eastern Suburbs line, Parramatta-Strathfield for T2 Inner West line, or Chatswood-North Sydney for T1 North Shore line) also see a surge in demand during this period while Sydney CBD is the busiest area with Central Station clocking up to serve an average of 95,426 passengers per day during peak hour. Nevertheless, the notion of PT density in off-peak hour is completely different from peak hour, which is illustrated as Figure 10 below.

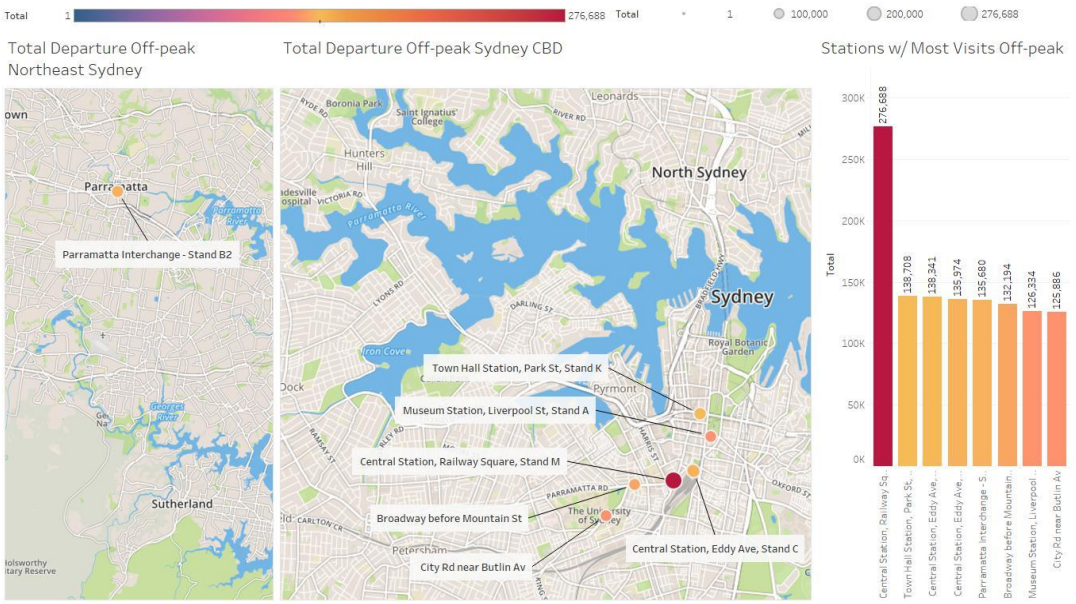

Figure 10 - Tap-on density during off-peak hour from 01 August 2016 to 19 September 2016

From Figure 10, it can be seen that the pressure on the station infrastructure during off-peak period is more loose compared to peak period. As well, the most popular transport mode during this period is surprisingly the bus, not the train, as the stations that serve the most visits outside peak hour are all bus stations; in particular, the noticeable detail occurring in Figure 10 is that 5 out of 8 stations with the most density are within the interchange point (Parramatta Interchange, Town Hall Station, and Central Station), where passengers change for connecting transport lines or to another transport mode. Correspondingly, it may be missing if the only focus is about the departure points of the PT commuters, so the tap-off density will also be investigated as illustrated in Figure 11 below.

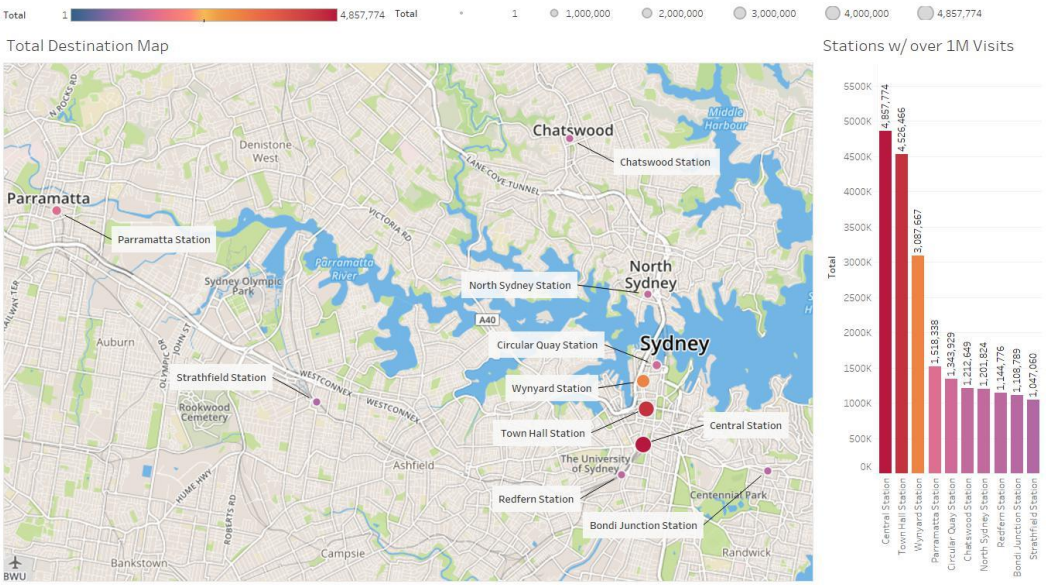

Figure 11 - Tap-off density from 01 August 2016 to 19 September 2016

As can be seen in Figure 11, the tap-off density during the observed period is similar to the tap-on density, which implicates that the above 10 stations serve most PT passengers as summarised in Table 6 below, from which it can be seen that out of 28,847 stations across NSW, the total trips made by passengers accessing the below 10 stations consists of nearly ¼ of the total trips, for both departure as well as destination. As well, Inner City stations are always crowded since they connect the PT lines from the northern suburbs with southern suburbs of Sydney, and they also offer all kinds of transport mode (train, bus, ferry, light rail) comparing to other stations which only offer bus or train (or both) service.

$N_{station}$ = 28,847	Departure	% of Total	Destination	% of Total
Central Station	4,771,297	5.32%	4,857,774	5.42%
Town Hall Station	4,541,360	5.07%	4,526,466	5.05%
Wynyard Station	2,899,769	3.24%	3,087,667	3.45%
Parramatta Station	1,459,307	1.63%	1,518,338	1.69%
Circular Quay Station	1,374,061	1.53%	1,343,929	1.50%
Chatswood Station	1,161,041	1.30%	1,212,649	1.35%
North Sydney Station	1,164,614	1.30%	1,201,824	1.34%
Redfern Station	1,068,111	1.19%	1,144,776	1.28%
Bondi Junction Station	1,122,970	1.25%	1,108,789	1.24%
Strathfield Station	1,068,854	1.19%	1,047,060	1.17%
TOTAL (N = 89,611,842)	**20,631,384**	**23.02%**	**21,050,272**	**23.49%**

Table 6 - Passenger visits summary of the 10 most crowded stations

Moreover, when examining the effect of off-peak and peak hour to the patronage of the PT network, it can be seen that the effect of interchange really has an impact on the density of patronage in PT network. As illustrated in Figure 12 below, the patronage density of tap-off activities concentrates mostly in the same stations as tap-on activities that has been mentioned, which means that these stations hold the significance of PT patronage.

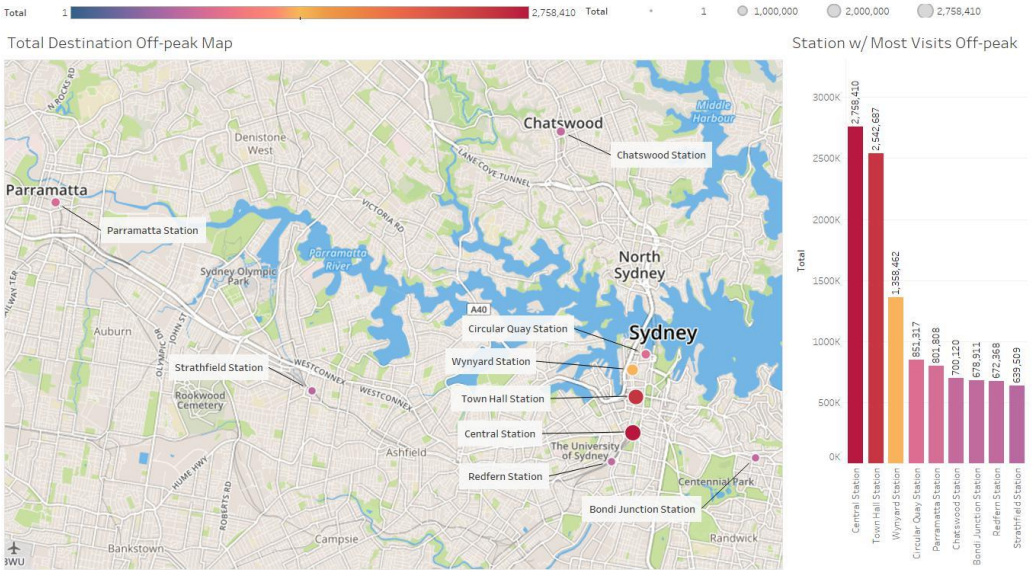

Figure 12 - Tap-off density during off-peak hour from 01 August 2016 to 19 September 2016

Furthermore, as can be seen in Figure 13, the stations with the most passenger tap-off activities during off-peak hour are somewhat the same as those with most passenger tap-off activities during peak hour, if not to mention that the volume of activities is only slightly different with off-peak activity volume being higher than peak activity.

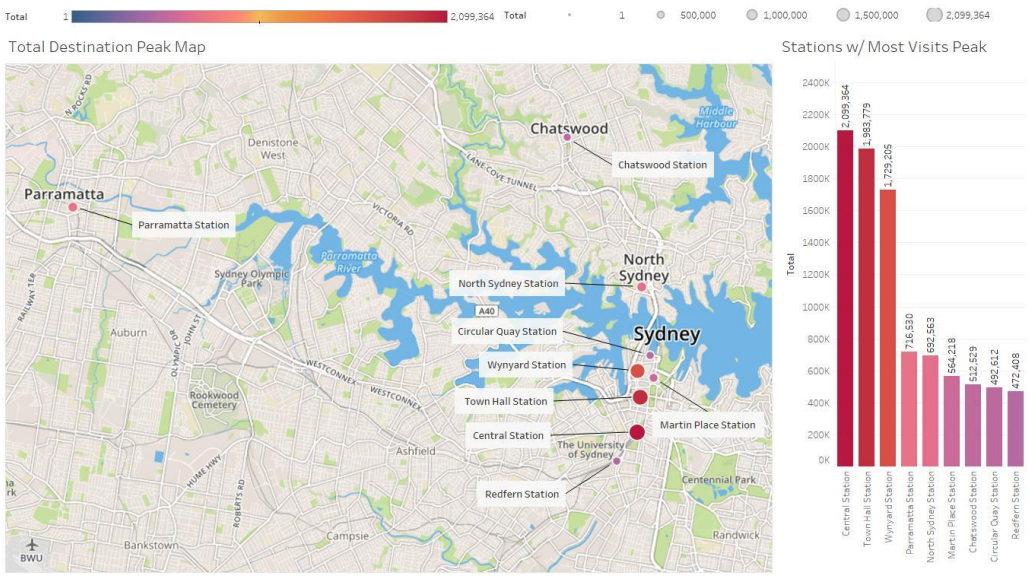

Figure 13 - Tap-off density during peak hour from 01 August 2016 to 19 September 2016

In summary, the analysis of valid Opal activities during peak and off-peak hours shows that there is a relationship between the PT mode shift and the PT line choice of the passengers since almost all the busiest stations shown above are interchangeable stations (except Bondi Junction Station) as can be referenced from the Sydney Trains Network map (available at website *https://transportnsw.info/document/2365/sydney-trains-network-map.pdf*). Furthermore, as analysed from Table 1, coupled with the transport patterns during peak and off-peak hours, it seems that even though

travelling during peak hour will cost them more money and time, passengers still choose to travel during peak hours for some reason; thus, further investigation into the passenger flow in the stations can unravel the true patterns of passengers during the application of peak hour.

As been concluded from the analysis of valid Opal activities during peak and off-peak hours, it can be seen that Sydney CBD stations are the intermediary of other surrounding stations since it offers all transit types as well as holds the largest interchange system in the NSW PT network (Central, Town Hall, and Wynyard Stations). However, some certain pattern can only be seen when examining deeper into the operation period of the station. Therefore, it is necessary to study about the density of commuters during each time period of the day. However, this study will be different from the previous study about overall density of each station since this study will include all instead of only valid tap-on and -off details. As such, Figure 14 results from querying the density of commuter visits of the 6 busiest stations in NSW PT network: Central, Town Hall, Wynyard, Circular Quay, Parramatta, Chatswood.

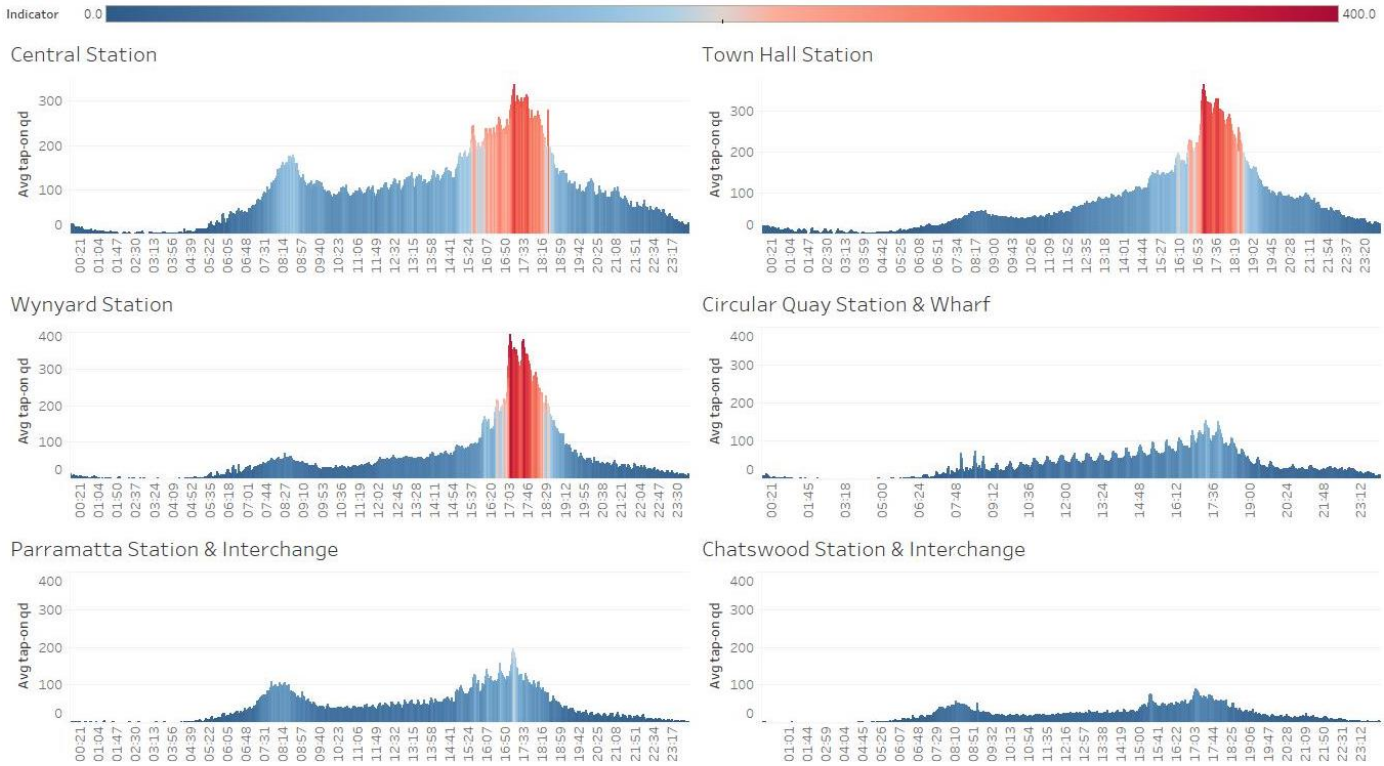

Figure 14 - Time density of average daily passenger visits per station

As can be seen from Figure 14, there is somewhat similar pattern derived from observing visit density of these 6 busiest stations. The lowest density that can be observed is after 10 PM until 4 AM the next day, this is the timeframe that most PT services have stopped operating or operate with lax schedule. The highest density that can be seen is during the period after 4 PM until 6:30 PM on the same day, this is the time that most people have finished their work and leave their offices to come back home. Another significant pattern that must be mentioned is the sudden increase in passenger visits from about 6:45 AM until 9 AM every morning, this is also the time that most people start coming to work so it will be a surge of passenger during this period. Outside of those important points, the passenger visits slowly decrease until mid-day before rising again just right after afternoon. Especially, in Table 7 below which is derived from querying the maximum and minimum density associated with each time point, it can be seen that the density at these busiest stations

reaches its climax during the first 20 minutes after 5 PM, the end of work day for most offices, while it reaches its lowest during the period after 1:30 AM to 3 AM, the time when all activities have long finished.

Stations (Average daily density)	Max density	Time point	Min Density	Time point
Central Station	337.3	17:12	0.3	03:02
Town Hall Station	364.5	17:09	0.1	02:50
Wynyard Station	394.6	17:07	0.1	02:43
Circular Quay Station and Wharf	155.3	17:19	0.1	02:18
Parramatta Station and Interchange	173.8	17:09	0.1	01:27
Chatswood Station and Interchange	89.76	17:07	0.02	01:32

Table 7 - Station density summary

Nevertheless, the elasticity of maximum density in Wynyard Station is a little bit narrower than other stations since it spans from 04:30 PM until 06:16 PM while other stations show signs of elasticity spanning from around 03:30 PM until 06:30 PM. Correspondingly, as shown in the following Figure 15, the same pattern of passenger visits per station is also applied for visits per transport mode, which signifies that the surge of passenger visits occurs during the start and the end of work day while the lowest density occurs by the end of the day until the start of new work day.

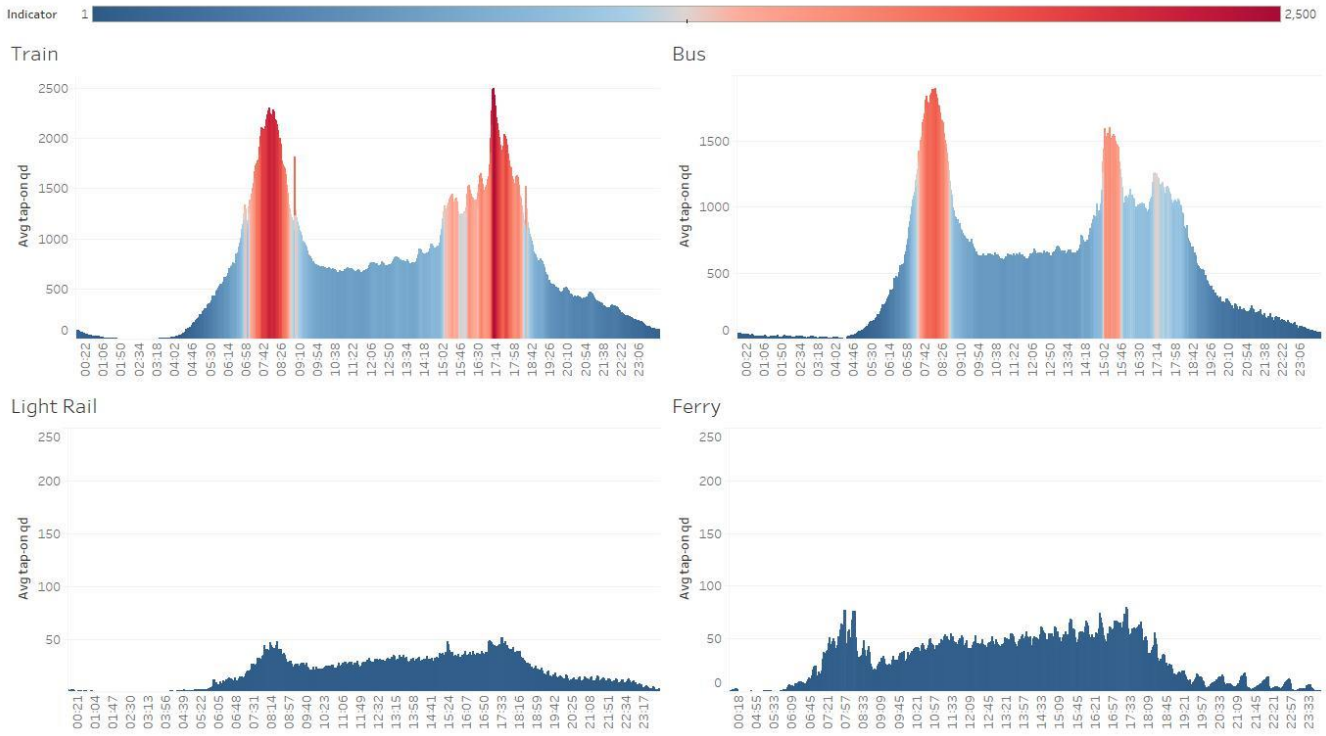

Figure 15 - Time density of average daily passenger visits per transit type

However, the pattern applied for transport mode is slightly different from one applied for stations as the scale of density for light rail and ferry is 10 times lower than train and bus, and the maximum density achieves in each transport mode is at different time, which is summarised as Table 8 and Table 9 below. From the figures shown in Table 8, it can be seen that the passenger patronage accessing the train network reaches its highest density during the first 10 minutes after work day ends while bus network achieves its highest density during the first 5 minutes after work day starts; as for light rail and ferry modes, since they make up for only a small amount of PT network, their density is not so high but they also somehow follow the same pattern as bus and train modes as shown in Figure 15.

Transit type (Average daily density)	Max density	Time point	Min Density	Time point
Train	2,500.00	17:09	1.00	02:24
Bus	1,904.00	08:04	2.00	02:24
Light rail	52.08	17.35	0.04	02:50
Ferry	79.54	19:25	0.02	00:25

Table 8 - Transport mode density summary

Transit type	First surge starts	First surge ends	Second surge starts	Second surge ends
Train	06:00	09:00	16:00	18:30
Bus	06:08	08:55	14:57	15:51
Light rail	07:45	09:04	16:43	18:18
Ferry	07:18	08:21	16:46	18:08

Table 9 - Transport mode density surge time

In addition, the information in Table 9 indicates that the density surge in train mode have most impact on the overall surge of passenger density in all stations since their patterns follows similar pattern as in train mode. However, it can also be concluded that commuters have tendency to use PT service before the start of morning peak hour which is signified in the start and end of first surge of passenger density in all 4 transit modes, while they mostly tend to use bus mode before the afternoon peak hour and use other 3 modes during afternoon peak hour.

In the same fashion, the inter-modal transfer journeys are drilled deeper, in which each transport mode will be presented with their busiest station during PT operation hour to further analyse the transport pattern of these transit modes. As can be seen in Figure 16, commuters tend to proceed to use train mode the most during the period from around 4 PM until 6:40 PM, which detailed information is summarised in Table 10 below.

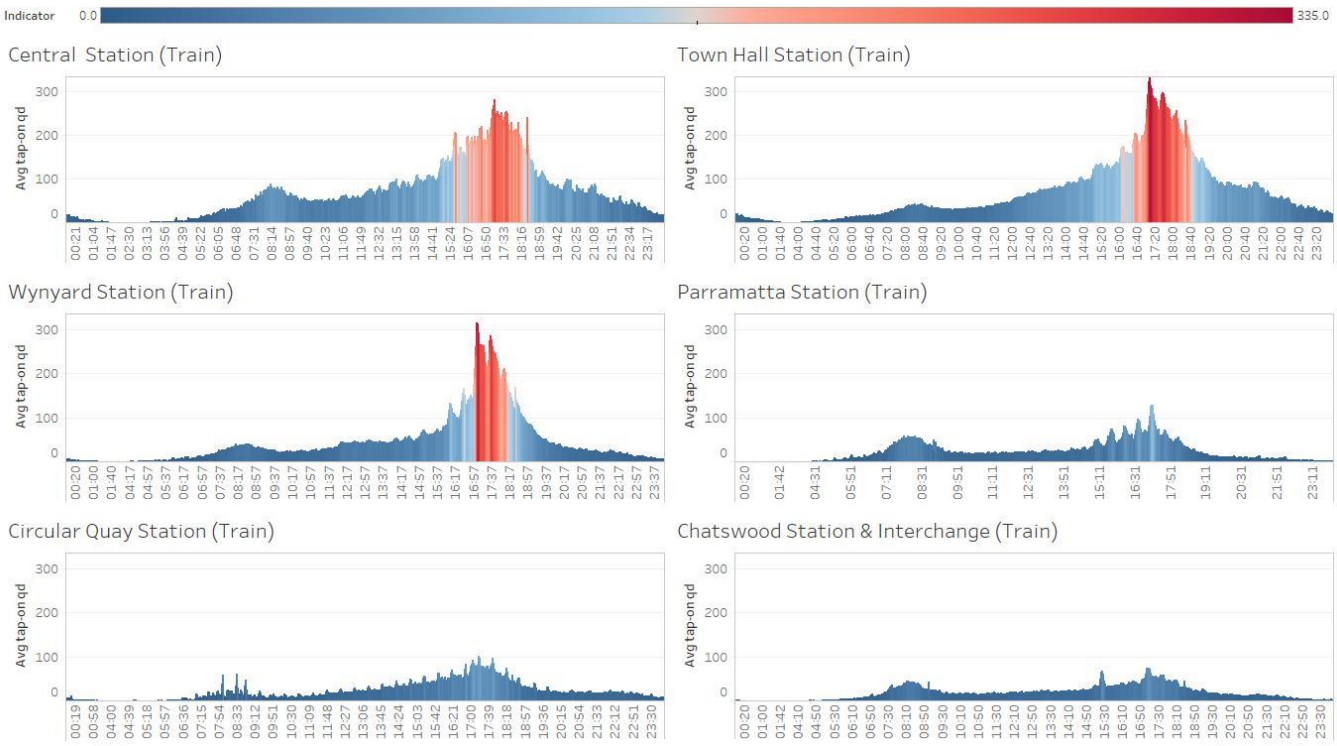

Figure 16 - Time density of average daily passenger visits through train mode per station

	Station (Avg density)	Max density	Surge start	Surge end
Train	Central	282.80	16:04	18:19
	Town Hall	332.10	16:31	18:42
	Wynyard	315.80	17:00	18:12
	Parramatta	128.90	17:00	17:31
	Circular Quay	100.10	16:56	17:53
	Chatswood	74.42	17:00	17:47

Table 10 - Train mode density surge time

From Table 10, it can be seen that, for train mode, the surge in density in Central Station starts earlier than other stations but lasts longer; however, the highest density belongs to Town Hall Station even though Central Station has the most customer served. As well, the inter-modal transfer journeys for bus are also illustrated and summarised as Figure 17 and Table 11 below.

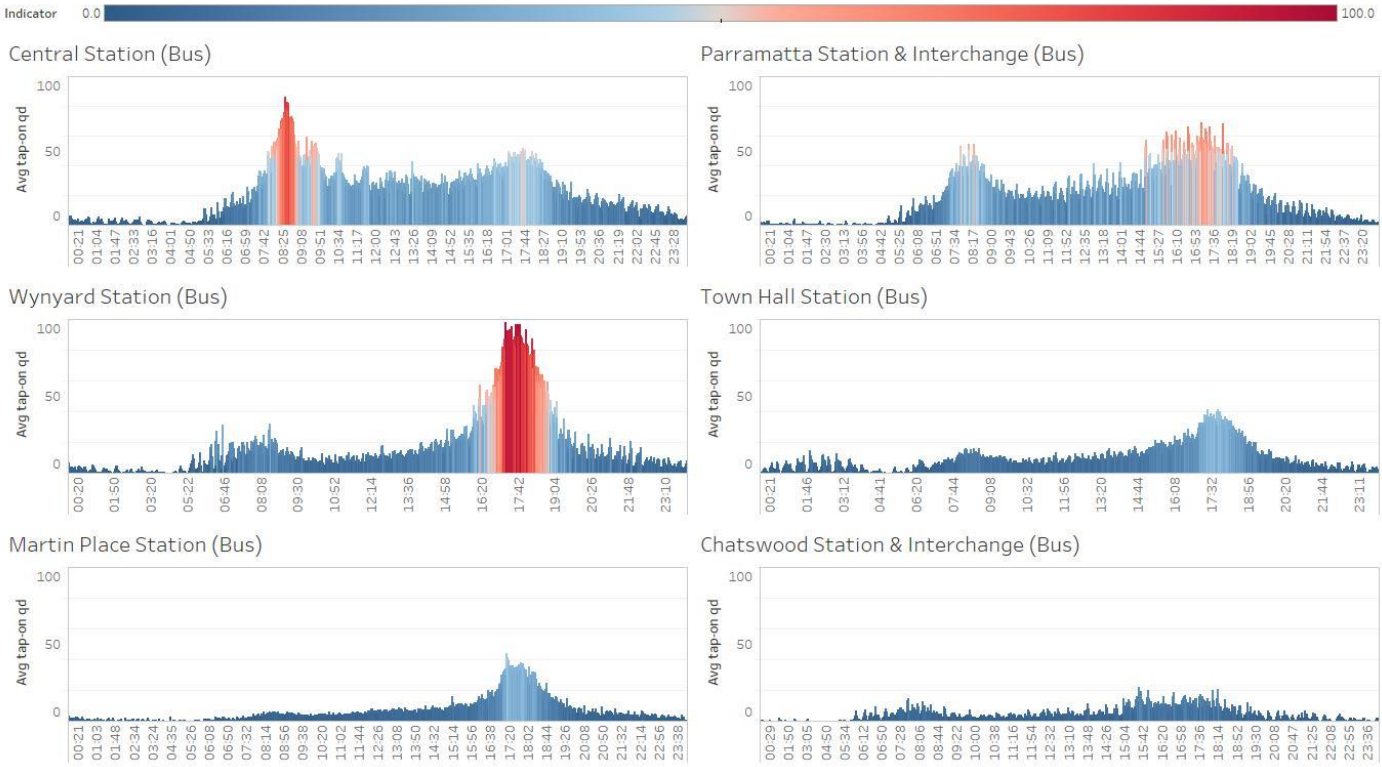

Figure 17 - Time density of average daily passenger visits through bus mode per station

	Station (Avg density)	Max density	Surge start	Surge end
Bus	Central	86.20	06:16	08:55
	Parramatta	68.66	15:38	18:19
	Wynyard	98.20	16:32	18:43
	Town Hall	41.34	17:00	18:25
	Martin Place	43.60	17:01	17:59
	Chatswood	22.18	15:28	15:55

Table 11 - Bus mode density surge time

From the above illustrations, the noticeable sign that can be observed is that the highest surge of customer happening in Central Station was during the morning when the end of peak hour was approaching while it happened in other stations during afternoon peak hour. Also, though the highest surge among the stations that can be observed belongs to Wynyard Station, the overall passenger flow during the day in Central Station seems to be more stable among observed stations.

Likewise, the travel patterns for ferry and light rail modes can also be uncovered using the same strategy, which can be derived from the following Figure 18 and Figure 19.

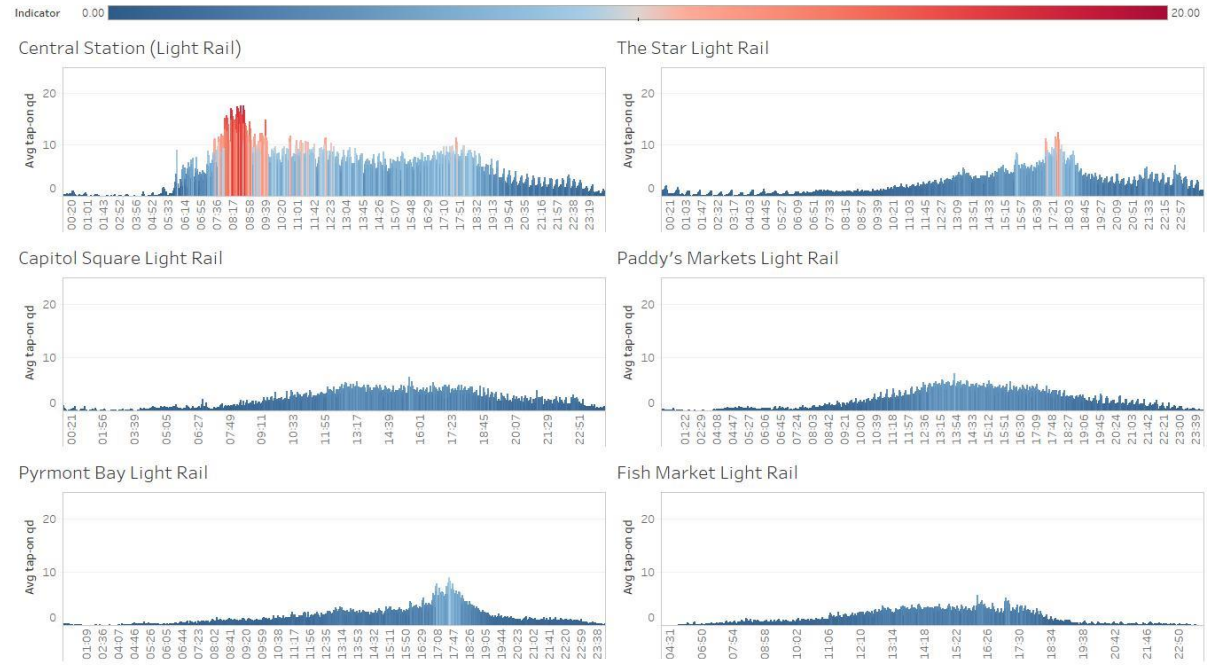

Figure 18 - Time density of average daily passenger visits through light rail mode per station

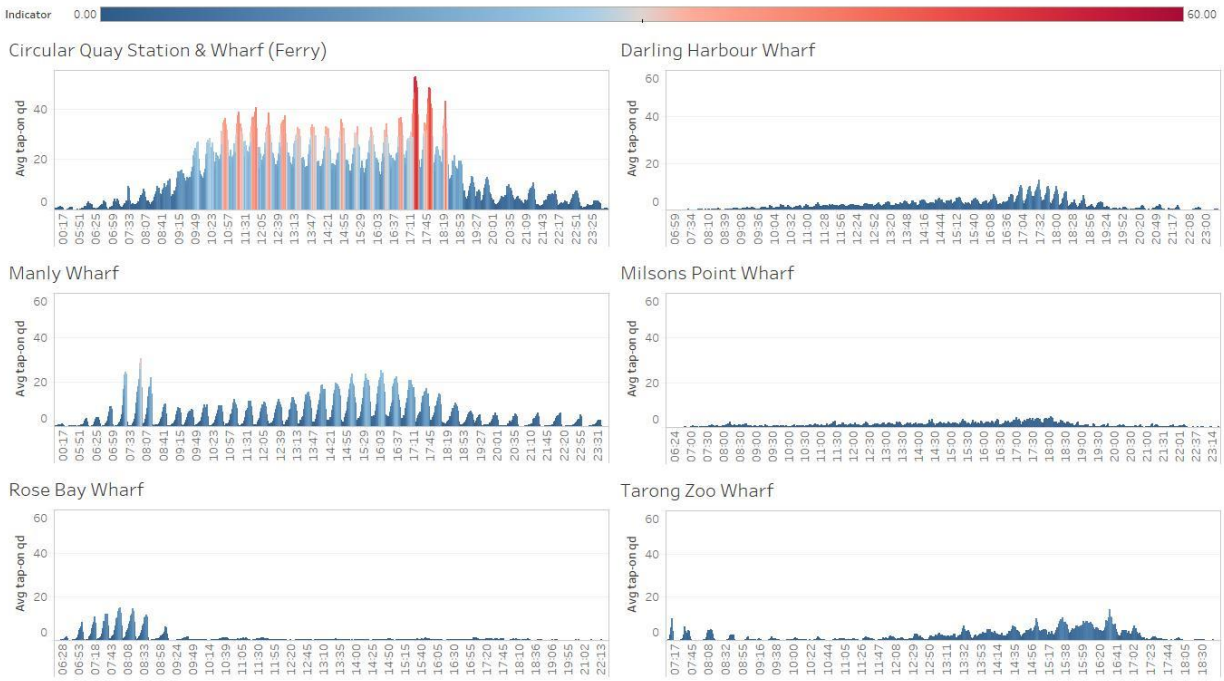

Figure 19 - Time density of average daily passenger visits through ferry mode per station

The pattern illustrated in Figure 19 shows no sign of recognisable pattern across ferry wharfs but there is one significant pattern entailing Circular Quay Wharf, which is that the demand for ferry across working hour is higher than other time of the day. This can be explained that ferry follows strictly the scheduled timetable and there is only the departure and destination between the start and the end of ferry route without any stops, so it is understandable that the pattern shown in the ferry activity looks more like "wavy" than "trendy". Per Figure 18, since light rail is only offered within Sydney CBD and its surrounding suburbs, the scope of this transport mode is a bit small and it only occupies a small amount of transport record; however, by observing it pattern, it can be seen that there is only one noticeable pattern that can be observed in the light rail mode is Central Station,

which shares the same pattern as its bus mode, while other stations appear to be insignificant and do not have any distinguished pattern to derive.

In essence, it can be concluded that passengers tend to start their trip during peak hour which results in the surge of demand from 6 AM to 7 AM in the morning and 3 PM to 4 PM in the afternoon as can be seen from the above patterns, this action can save them 30% off their fare. Moreover, although off-peak hour discount is only applied for train mode, the trend of passengers in train mode shows that the density of the surge in demand for train mode during the time from 6 AM to 7 AM is very high even though it does not reach the climax yet, which means that the passengers really do consider their mode choice based on the cost effectiveness if it is within their ability. One more point that can be made from this analysis is that the mode choice relies heavily on the accessibility that the stations offer. This claim can be ascertained through the balance of demands made for stations with highly interchangeable ability such as Central Station which connects all five major PT railway lines and also offer wide range of other transport modes (light rail and bus).

Last but not least, the effect of interchange must also be accounted for when analysing the travel pattern of PT users. For starters, Figure 20 showcases the summary of inter-modal transfer made during the observed period, which highlights that the transfers between train and bus are the most popular type of inter-modal transfer. This pattern is understandable since train and bus make up most of PT routes in NSW PT network so the interchange between these 2 models must be the highest. Furthermore, Central Station is the interchange point that has the highest volume of inter-modal transfer made.

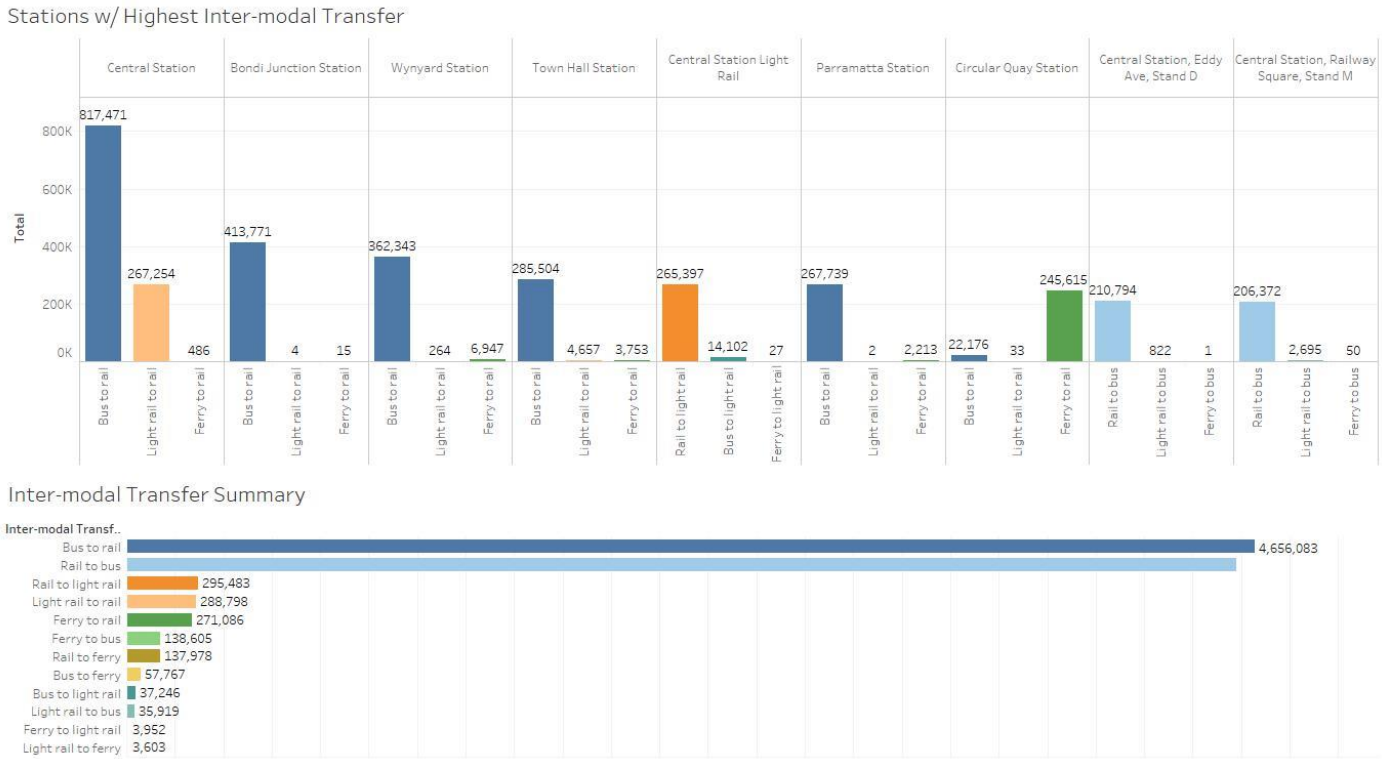

Figure 20 - Inter-modal summary

After checking the activity of inter-modal transfer during off-peak and peak hour, some interesting patterns has emerged, which is illustrated in Figure 21 and 22 below. As per Figure 21, it can be seen that the focus of mode shift during the off-peak hour is the bus since almost every transfer is from

another transport mode to bus, there are some mode shift to ferry, but the amount of those transfers is insignificant when comparing to the transfers to bus.

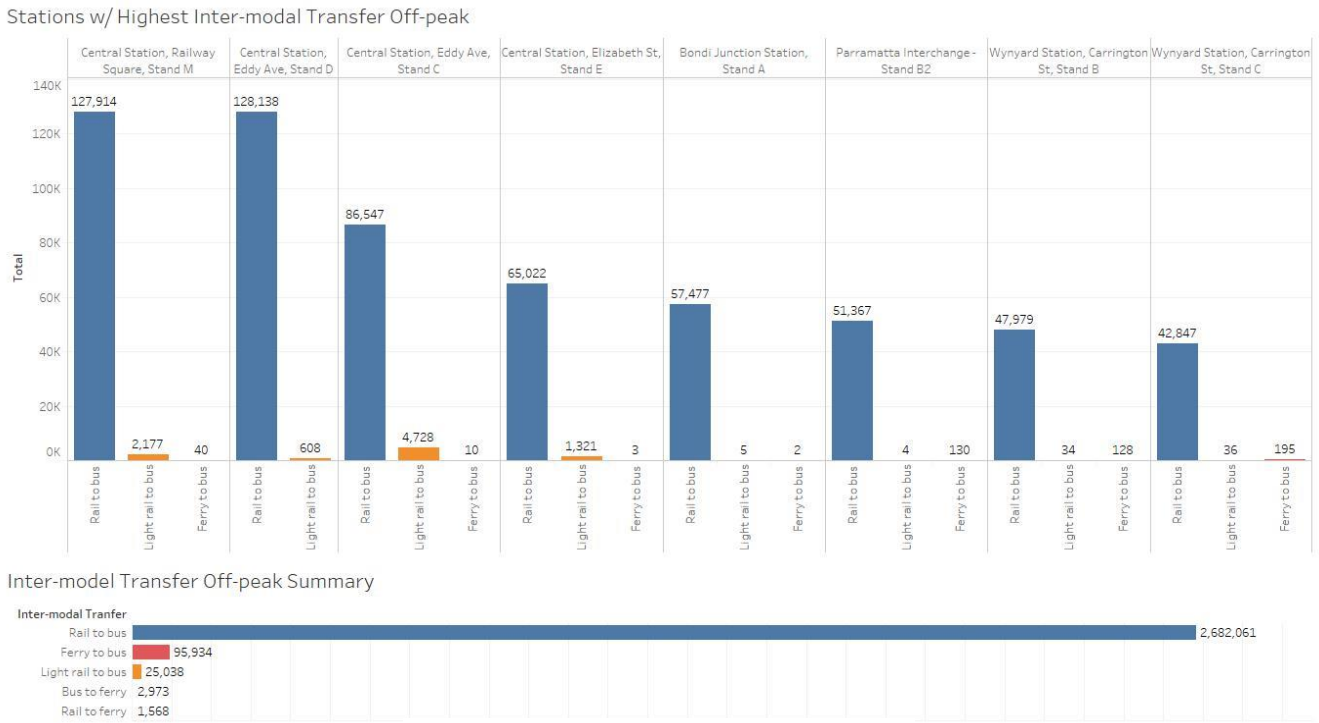

Figure 22 - Inter-modal transfer summary off-peak

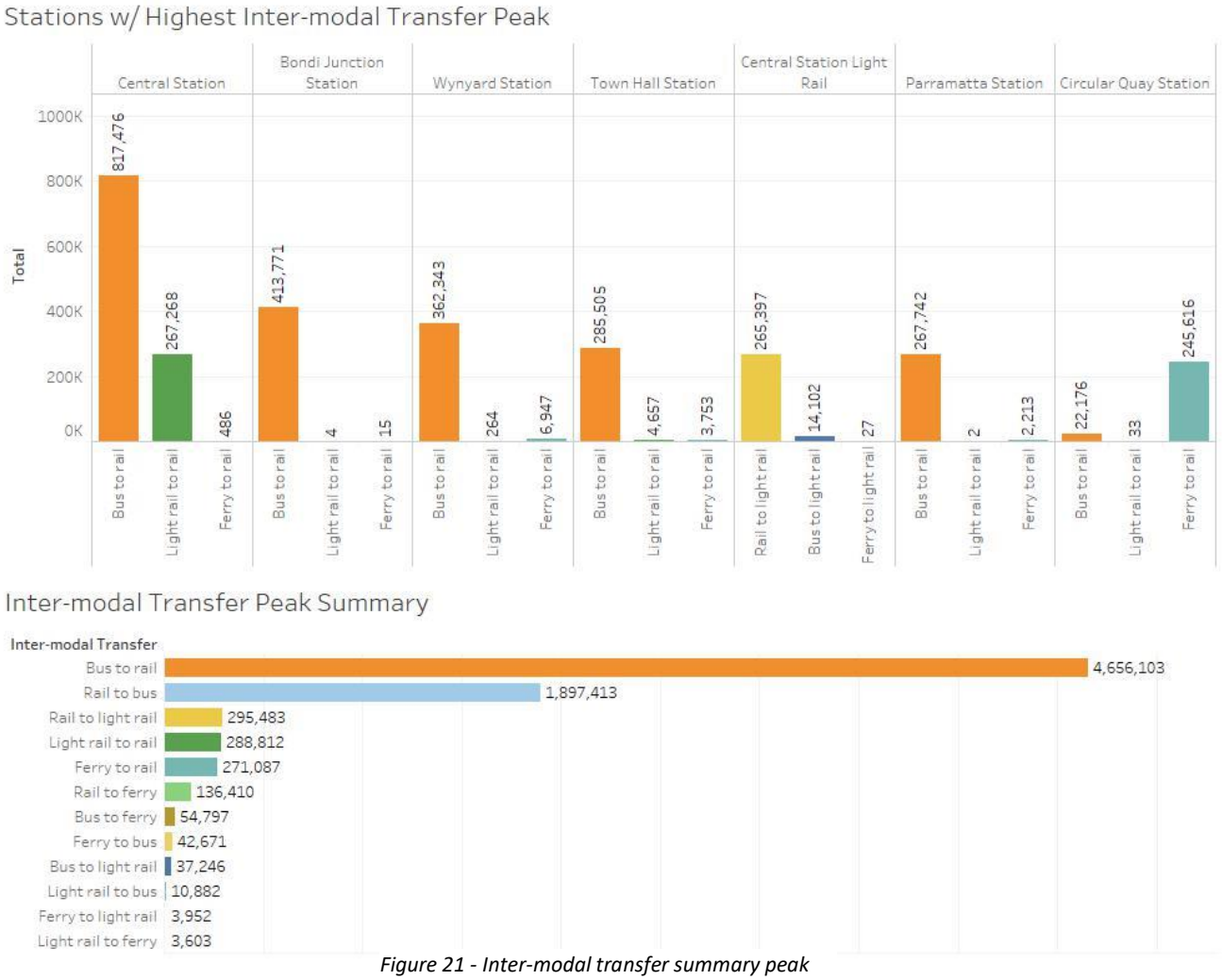

Figure 21 - Inter-modal transfer summary peak

As for Figure 22, it shows that the mode shift during peak hour mostly involves bus, train and light rail, while a small potion has the appearance of ferry mode. Through the combination of the effect of mode shift during off-peak and peak hour, it is believed that PT users pattern changes based on the significance of the application of off-peak hour since their behaviours are totally different when comparing between peak and off-peak hour. The last item that needs investigating is the habit of inter-modal transfer among PT users, which is highlighted in Figure 23.

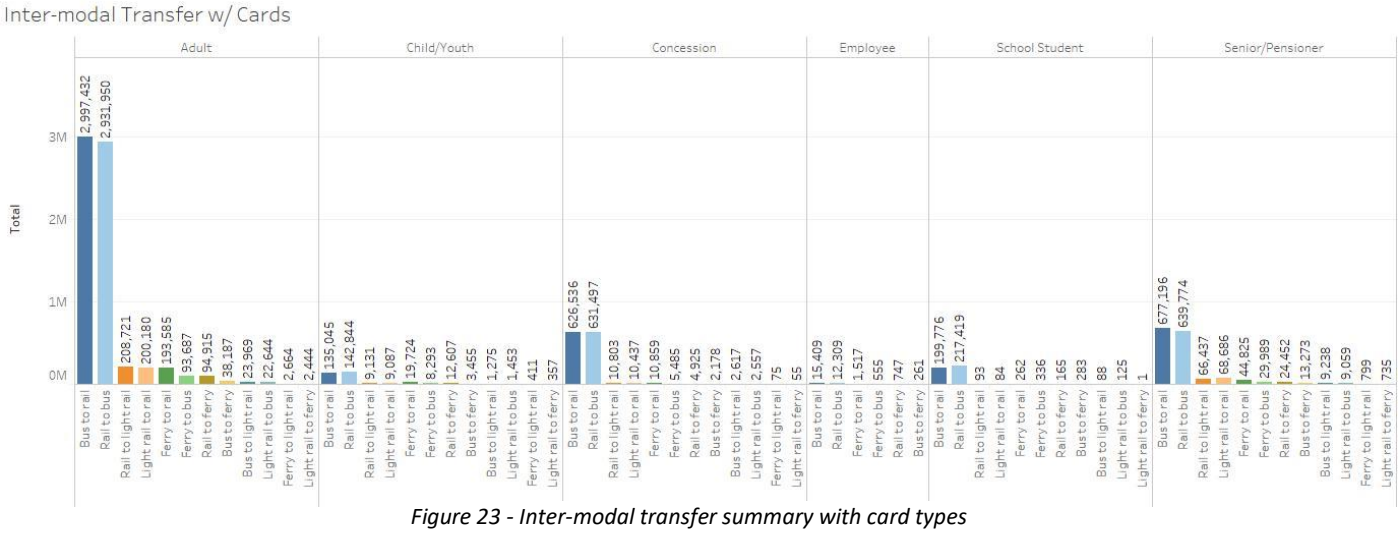

Figure 23 - Inter-modal transfer summary with card types

As can be seen from the above figure, most inter-modal transfers made by each type of Opal card user involve in the exchange between train and bus. Light rail may play some parts in shaping mode shift patterns of PT users, however, not all areas offer light rail service, so its effect is limited. This situation also applies for ferry since ferries only exist where the wharfs are built, so their effect to mode shift on PT users is also limited. When taking the closer look into the behaviour of each type of card holder, it can be seen that their behaviours can be distinguished easily as shown in Figure 24 below.

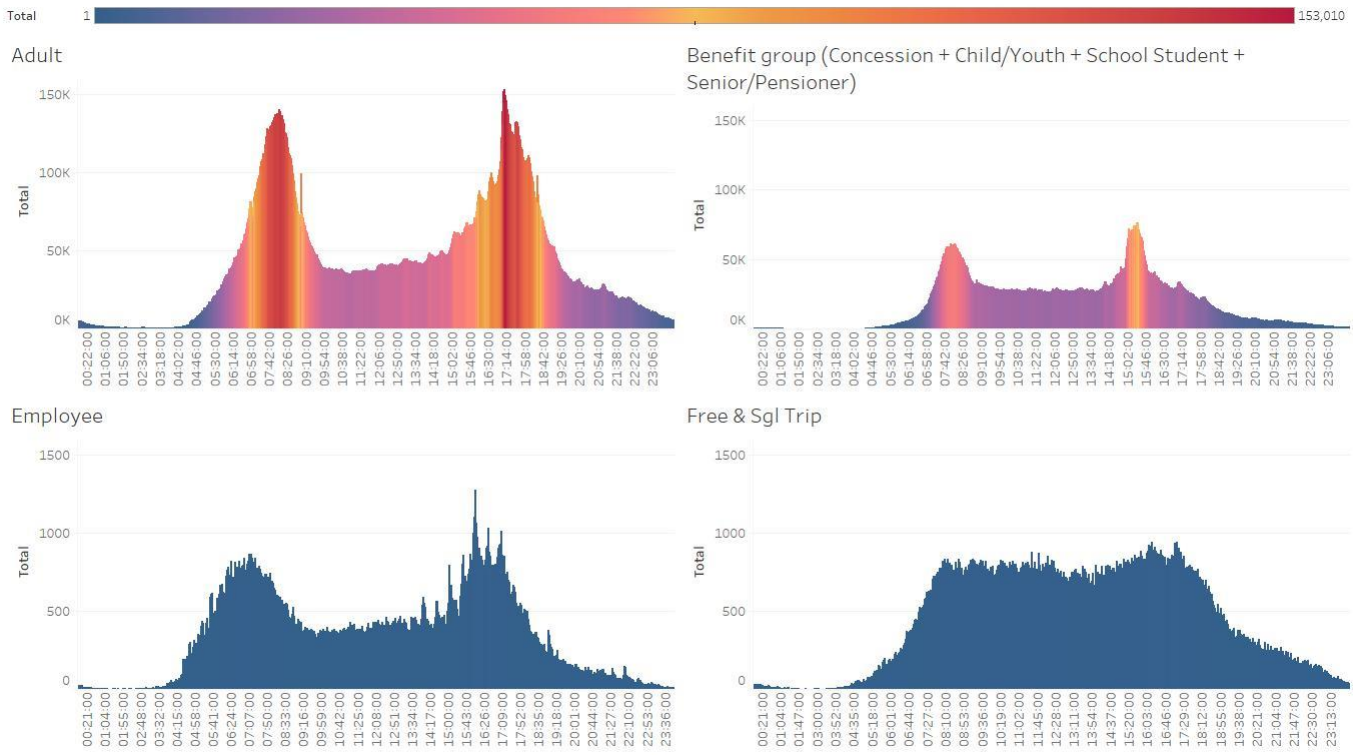

Figure 24 - Timely distribution of each card type activity

As can be seen from Figure 24, the most significant sign of peak hour avoidance can be seen within the group of Employee card type, in which the surge of tap-on event starts from around 5 AM and reaches it climax at around 7 AM before gradually decreasing until returning to normal at around 9 AM. This pattern has shown that the group of employee has clear intention to avoid peak hour tap-on so that they can enjoy 30% off-peak discount. This pattern can also be seen right before afternoon peak hour though not really as clearly as in the morning peak hour. As well, the same pattern can also be seen through the Benefit group, though it starts later during the period before morning peak hour; however, it can be seen clearly during the period before afternoon peak hour as it starts by around 2 PM and reaches climax around 3:15 PM. This can be explained that the Benefit group consists of school students and children or youths, since their school day usually finishes by around 3 PM so it is normal that they usually tap-on during this time to travel back home. However, there is no significant sign in the Free and Single Trip group since they are free trips or flat rate trips, which are not affected by off-peak discount benefits. Meanwhile, the Adult group also show signs that they also have tendencies to avoid peak hour fare, but this tendency is weaker than the Employee or Benefit group, which starts at around 6 AM and reaches climax at around 7:45 AM in the morning and starts at around 2:20 PM and reaches climax at around 5:10 PM. Therefore, it can be seen that a minority of Adult group who travels earlier during the surge can enjoy the off-peak discount if they travel by train.

In conclusion, the activities and interchange of bus and train have the most effect on the mode shift of PT users, which contributes to the fact that their behaviours and travel patterns are recognised through the analysis of their activities through off-peak and peak hours of the PT network which is combined to become their mode choice. It is believed that passengers have tendencies to choose to travel by train right after peak hour starts, which results in the sudden increase in tap-on events during the 30 minutes interval before the start of peak hour. However, analysis from the mode shift summary has shown that they mostly change from bus to train during peak hour and train to bus during off-peak hour, so there may be some underlying meaning behind this action. Upon checking the surge of patronage which has been presented, it is confirmed that the overwhelming amount of mode shift from bus to train results from the behaviour of PT user by the end of work day when they do not care about off-peak hour discount and proceed right onto train to travel back (which can be seen in Figure 13 where most tap-on activities for train mode occurred during peak hour from 4 PM to 6:30 PM).

CHAPTER 4
DISCUSSION

The result from the quantitative and qualitative research shows that transport pattern is decided based on the habit of travelling of the users, the available transport mode that is offered in the stations, the effect of off-peak/peak hour application, the accessibility to the station, and the infrastructure and service quality that the station can offer to its passengers. The result also reveals that despite the fact that travelling by train during peak hour may cost commuters 30% more than they usually pay during off-peak hour, the majority of train trips made during the observed time frame stem from the peak hour activity. However, the benefit of off-peak hour discount can still be seen from the result of this report due to the fact that even though the number of off-peak hour trips is inferior to the number of peak hour trips, there is still an undeniable fact that the average fare spent on off-peak hour trips is 34% less than average fare spent on peak hour trip, which is further supported by the policy of off-peak hour discount fare, where travelling during this time cost passengers 30% less than during peak hour. As well, the perception level of satisfaction of the PT users is enhanced proportionally to their time spent in utilising PT service due to the fact that there may be some occurrences that may not be in favour to the PT users when they use PT service; however, by interacting with the service long enough, their perception will gradually become more precise and their judgement on PT service will become more keen. Furthermore, PT users have not been totally introduced about the benefits of acquiring an Opal account in using NSW PT system, which is ascertained by the fact that the growth of single trip tickets over the observed period is very high, and they even precede the total number of Opal accounts of some card types. Additionally, the overcrowding factor in NSW PT network has gone to an alert level since travelling during peak hour has been proved during the procedure of this report that it takes passengers 2 times longer the time when comparing to travelling during off-peak hour. Lastly, the most important notation that can be taken from the procedure of this project is that the data collection procedure of NSW PT system is facing some problems since the ratio of data collection error is very high, only an in-depth analysis on data structure as well as applying complex query functions can temporary solve this problem.

As can be seen, accessibility in PT is one of the most significant factors which denotes the performance level of transit system. Specially, for regular commuters who do not have their private vehicle as well as customers with limited mobility, PT accessibility is the determinant factor and remarkable quality related features for their decision for route preference, usage, mode choice and departure time selection. The improvement of PT accessibility may encourage the private vehicle users to change the travel mode to PT and could help for the environmental concerns as well. If the primary purpose of transit-oriented design is creating land-use pattern which make transit accessible for potential riders, then for this situation, transit access must be situated in the rider's current place and the destination. Hence, biking and walking accessibility to PT could be the important concerns for integrating land use for the transit planning. This factor could influence the users' perception degree and overall transit systems satisfaction. Many researchers have indicated that walking is the most significant and natural mode to access public transit. Thus, walking accessibility to PT is the indicator of performance or quality of PT service (Mintesnot and Rita 2011).

Moreover, some of the factors such as attractiveness of environment and aesthetic quality, street connectivity and infrastructure provision also have impact for the ease of access to PT. Walkable environment could be improved by establishing pedestrian supporting infrastructure, high-quality architectural design and visual amenity, permeability and street connectivity. One of the researchers states that suitable environment should be provided for encouraging the commuters to walk and for using public transport. Furthermore, for the efficient PT network, those stated environments should ensure the safety of the commuters and should be interesting, stimulating and comfortable. Neighbourhoods with integrated and permeable road network can provide multiple direct route options for both services and pedestrians, and hence encourages many people to use PT. The architecture and appearance of buildings also help to increase the attractiveness and aesthetic quality of an environment which could attract many commuters to take the service of public transport. Public frontages, different visual architecture and streets having activities also help the travellers to feel more comfort, secure and interesting. Some of the facilities such as sufficient public spaces and footpaths with sight distance could help to feel the customer more secure (Chowdhury, Zhai and Khan 2016).

As well, there are so many benefits of improved transport system, but travel time reliability gains and travel time savings are two basic factors which affect the quality of public transport service. Generally, appropriate assessment of any type of transport system needs monetary judgment for both the value of travel time reliability (VOR) as well as the value of travel time (VOT). This should be considered by the policy makers who has an option to choose public transport infrastructure projects having reliability gains (for example: constructing the bypass having more capacity) or/and travel time savings (for example: constructing high speed transport mode) (Beaud, Blayac and Stephan 2016).

According to Dong and Yan (2015), the two most significant barriers that inhibit the mode choice for PT users are that it takes them too long to travel to the destination, and the cost considerations may force them back. As well, they discuss about the two most important facilitators on encouraging mode choice which are the comfort during the trip and reasonable fares for the trip. As per analysed on the density of each PT station, there is evidence that the PT vehicles may encounter overcrowding situation at some point, especially during peak hour, which can be seen with a surge of passengers travelling in transport modes. Furthermore, there is also evidence that the surge event takes place for quite a period before showing sign of cooling down. As such, the prolonged crowded period will gradually increase the time for current passenger to hop off the train and new passengers to hop on the train as train driver has to wait until being given a sign to start heading towards the next stop. Therefore, the prolonged time will create 1 of the following 2 scenarios:

- The train will be late for the next stop; or,
- the timetable will be prolonged during peak hour to adapt to the change.

Either of these 2 scenarios is not advantageous for the passengers since prolonged waiting time will see the number of customers increasing minute by minute, then the situation will become worse and worse until the peak period cools down. Besides, when considering about the PT fare during peak hour, the passengers will also have 2 options:

- They will wait until the peak hour cools down, then take a transport to travel back; or,

- they will afford a peak fare rate at 30% more than off-peak fare if they pick the former option.

Either way, these 2 options do not also seem really good for the passengers since they have to delay the time for travelling back home or have to spend more than they usually spend. As supported from Alejandro, David and John (2013), they discuss that the crowding event in PT systems may have bad effect on waiting time, travel time reliability and savings, route choice, as well as optimal fare. Accordingly, they argue that due to limited size of PT vehicles, there may not have enough comfortable space for passengers if a large number of new passengers must be handled. This situation is somewhat similar to Japan whether the PT passengers must stay squashed inside the carriage during peak hour just to travel from home to work/school and vice versa. They also argue that overcrowding may also affect negatively to the mental wellbeing of the passengers where they do not only feel uncomfortable being squashed inside a small area but also feel anxious due to negative mental impact.

Nevertheless, the research conducted on this project is still lacking in terms of scale and time. Further research on this project topic should not only focus on static data of the tap activities of PT users but also give attention another kind of data that closely related to the former: timetable schedule. By coupling the collected data from PT users and the timetable based on their actual schedule and planned schedule, researchers may be able to discover what really happens at each station points of the PT network. For example, even though a train or bus arrive their destination on time, there is no certain support to be sure that they arrive at the stops on time as they may arrive at some check points earlier than schedule and pick up the passengers, or they may arrive late at some check points due to heavy traffic, or they may not stop at the check point and head straight to the next check point due to the fact that no one has reached their destination or no one is waiting to depart on the vehicle. As can be seen, the scenario happening during the operation of a PT route may be varied, and the scenarios do not repeat themselves, so it may be better if the timetable with their planned schedule as well actual activities can be consolidated with the tapping data of PT users. However, this action may make the volume of the dataset become huge; so, future research may also aim to provide an abstract data structure from this consolidation, which is not only reliable enough to represent the actual patterns or activities of the PT network, but also compact enough so that usual machines can process it as the procedure of big data analysis of this project always face with 2 problems of Hadoop ecosystem: Java heap size and data spill. These errors signify that the memory of the virtual machine cannot handle the amount of data anymore and need restarting to refresh the memory. By all means, it is undeniable fact that the study conducted in this project is only the tip of the iceberg and there are so many things to do to enhance its outcome, given that the allotted time allows. Another matter that needs handling is that the collected dataset contains majority of unknown data records or unrecognisable data records, which returns to be useful after reconstruction of dataset. However, there is still around 5.19% data records that must be removed due to being unable to be recognised; therefore, necessary parameters must be taken in order to improve data quality so that the percentage of unrecognisable data records can be reduced as minimal as possible. By improving the quality of data, the collected dataset will become more reliable and will not cost much time for reconstruction.

CONCLUSION

In this project, the qualitative research was conducted in order to capture preliminary opinions of PT users so that some basic ideas will be deducted before starting to design the quantitative research based on the de-identified Opal card data for discovering PT users travelling pattern. As the result of this project implies, transport pattern is decided based on the habit of travelling of the users, the available transport mode that is offered in the stations, the effect of off-peak/peak hour application, the accessibility to the station, and the infrastructure and service quality that the station can offer to its passengers. It is suggested that the status of timetable schedule (planned and actual) should be integrated into the data so that data analysis can address in more details about the problems occurring during the operation of PT network in order to plan for amendment of the problem. However, an unrelated problem emerges during the procedure of the project that the data collection of the dataset has encountered many data errors that needs fixing, which implies that in order to improve the insights from analytics in the future, it is a must to improve data quality. The result of this project suggests that in order to improve PT service quality, further system evaluation, accessibility, as well as perceived value improvement must be conducted so that more users will become interested in NSW PT network.

REFERENCES

Alejandro T., David A. H., John M. R. 2013, 'Crowding in Public Transport Systems: Effects on Users, Operation and Implications for the Estimation of Demand', *Transportation Research Part A*, vol. 53, pp. 36-52.

Andrew S. 2018, 'Groups Unite for Public Transport Rally', *Green Left Weekly*, 6 February, pp. 6.

Australian Bureau of Statistics (ABS) 2008, *Australian Social Trends*, cat no. 4102.0, viewed 22 August 2018, <http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4102.0Chapter10102008>.

Australian Bureau of Statistics (ABS) 2014, *Age Standard*, cat no. 1200.0.55.006, viewed 23 August 2018, <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/1200.0.55.006main+features62014,%20Version%201.7>.

Beaud M., Blayac T. & Stephan M. 2016, 'The Impact of Travel Time Variability and Travellers' Risk Attitudes on the Values of Time and Reliability', *Transportation Research Part B: Methodological*, vol. 93, pp. 207-224.

Becky P. Y. L., Cynthia C. & Eric T. H. C. 2010, 'Rail-based Transit-oriented Development: Lessons from New York City and Hong Kong', *Landscape and Urban Planning*, vol. 97, pp. 202-212.

Bruno A., Catherine M., Martin T. 2006, 'Mining Public Transport User Behaviour from Smart Card Data', *IFAC Proceedings Volumes*, vol. 39, no. 3, pp. 399-404.

Bureau of Transport Statistics (BTS) 2016, NSW and Sydney Transport Facts, NSW Government Canberra, viewed 04 August 2018, <https://www.transport.nsw.gov.au/sites/default/files/media/documents/2017/NSW%20and%20Sydney%20Transport%20Facts%202016.pdf>.

Calimente J. 2012, 'Rail Integrated Communities in Tokyo', *Journal of Transport and Land Use*, vol. 5, no. 1, pp. 19-32.

Chowdhury S., Zhai K. & Khan A. 2016, 'The Effects of Access and Accessibility on Public Transport Users' Attitudes', *Journal of Public Transportation*, vol. 19, no. 1, p. 97-113.

Craig M., Brian C. & Jillian A. 2016, 'Customer Perceptions of Quality of Service in Public Transport: Evidence for Bus Transit in Scotland', *Case Studies on Transport Policy*, vol. 4, pp. 199-207.

Dimitrios E. & Constantinos A. 2017, 'Understanding the Effects of Economic Crisis on Public Transport Users' Satisfaction and Demand', *Transport Policy*, vol. 53, pp. 89-97.

Dick M. 1998, 'Web-Based Market Research: The Dawning of a New Age', *Direct Marketing*, vol. 61, no. 8, pp. 36-38.

Dong W & Yan L. 2015, 'Factors Influencing Public Transport Use: A Study of University Commuters' Travel and Mode Choice Behaviours', *State of Australian Cities Conference 2015*, pp. 1-14.

Richard B. E., Adrian B. E., Stephen P. G. & Breno S. 2017, 'Electronic Ticketing Systems as A Mechanism for Travel Behaviour Change? Evidence from Sydney's Opal card', *Transportation Research Part A: Policy and Practice*, vol. 99, pp. 80-93.

Girts B. 2014, 'Commuting Patterns in Riga Agglomeration: Evidence from a Survey Analysis of Youth', *Regional Formation & Development Studies*, vol. 14, pp. 16-29.

Grosvenor T. 2000, 'Qualitative Research in the Transport Sector', *Transportation Research Circular E-C008: Transport Surveys: Raising the Standard*, viewed 02 September 2018, <https://pdfs.semanticscholar.org/8b9b/107cf9209ea631d811455f9331cf14ea7cff.pdf?_ga=2.192013805.694112197.1537885502-1340490483.1535857965 >.

John D. N. & Corrine M. 2013, 'The Impact of The Application of New Technology on Public Transport Service Provision and the Passenger Experience: A Focus on Implementation in Australia', *Research in Transportation Economics*, vol. 39, no. 1, pp. 300-308.

Kahle L. R. 1997, 'The Real-Time Response Survey in New Product Research: It's about Time', *The Journal of Consumer Marketing*, vol. 14, no. 3, pp. 234-248.

Kelly J. C. & Susan L. H. 2003, 'Qualitative Methods in Travel Behaviour Research', *Transport Survey Quality and Innovation*, pp. 283-302.

Lauren R., Margareta F., Tommy G. & Terry H. 2013, 'Quality Attributes of Public Transport that Attract Car Users: A Research Review', *Transport Policy*, vol. 25, pp. 119-127.

Li M., Nicholas H. & Michael T. 2011, 'Increasing the Patronage of Adelaide's Northern Rail Corridor', *The 34th Australasian Transport Research Forum (ATRF) Proceedings*, pp. 1-16.

Martijin B., Moshe G. & Piet R. 2009, 'Access to Railway Stations and Its Potential in Increasing Rail Use', *Transportation Research Part A: Policy and Practice*, vol. 43, no. 2, pp. 136-149.

Matt O. 2017, 'Overcrowding on Sydney's Trains Rapidly Getting Worse as Demand Soars', *The Sydney Morning Herald*, 15 December, viewed 20 August 2018, <https://www.smh.com.au/national/nsw/overcrowding-on-sydneys-trains-rapidly-getting-worse-as-demand-soars-20171215-h053rm.html>.

McLeod S., Scheurer J. & Curtis C. 2017, 'Urban Public Transport: Planning Principles and Emerging Practice', *Journal of Planning Literature*, vol. 32, no. 3, pp. 223-239.

NSW Independent Pricing and Regulatory Tribunal (NSW IPART) n.d., *Weekday Peak and Off-Peak Fares*, NSW Independent Pricing and Regulatory Tribunal, viewed 20 August 2018, <https://www.ipart.nsw.gov.au/files/sharedassets/website/trimholdingbay/information_paper_2_-_weekday_peak_and_offpeak_fares.pdf>.

Norman A. 2011, 'A Hard Reign: NSW Public Administration under Labor - 2007 to 2011', *Australasian Parliamentary Review*, vol. 26, no. 1, pp. 41-52.

Robert F. M. 2009, *Thematic History of the NSW Railways*, Office of Rail Heritage (RailCorp), Sydney, NSW, Australia.

Robert W. E. 2015, 'Convenience Sampling, Random Sampling, and Snowball Sampling: How Does Sampling Affect the Validity of Research?', *Journal of Visual Impairment & Blindness*, vol. 109, no. 2, pp. 164-168.

Shannon R. & Christina D. B. 2018, 'Focusing on the Fundamentals: A Simplistic Differentiation Between Qualitative and Quantitative Research', *Nephrology Nursing Journal*, vol. 45, no. 2, pp. 209-212.

Tommy G. & Geertje S. 2007, 'Travel Demand Management Targeting Reduced Private Car Use: Effectiveness, Public Acceptability and Political Feasibility', *Journal of Social Issues*, vol. 63, no. 1, pp. 139-153.

Todd L. 2017, *Understanding Transport Demands and Elasticities: How Prices and Other Factors Affect Travel Behaviour*, Victoria Transport Policy Institute, viewed 15 October 2018, <http:// www.vtpi.org/elasticities.pdf>.

Transport for NSW (TfNSW) n.d., *History of the NSW Railways*, Transport for NSW, viewed 05 August 2018, <https://www.transport.nsw.gov.au/projects/community-engagement/sydney-trains-community/culture-and-heritage/history-of-nsw-railways>.

Wen-Tai L. and Ching-Fu C. 2011, 'Behavioural Intentions of Public Transit Passengers - The Roles of Service Quality, Perceived Value, Satisfaction and Involvement', *Transport Policy*, vol. 18, pp. 318-325.

William G. C. 1977, *Sampling Techniques*, 3rd edn, John Wiley & Sons, New York.

Mintesnot G. W. & Rita C. 2011, 'Factors Affecting Traveller's Satisfaction with Accessibility to Public Transportation', *European Transport Conference*, pp. 1-10.

Yulin L. & Phil C. 2013, 'Spreading Peak Demand for Urban Rail Transit through Differential Fare Policy: A Review of Empirical Evidence', *Australasian Transport Research Forum 2013 Proceedings*, pp. 1-35.

APPENDIX A
QUALITATIVE DESIGN QUESTIONNAIRE

❖ Question 1: What is your gender?

a) Male b) Female

❖ Question 2: What age group do you belong to?

a) Under 18 b) 19-25 c) 26-35 d) 36-50 e) 51-70 f) Over 70

❖ Question 3: What is your occupation?

a) Self-employed b) Employed Full Time c) Employed Part Time d) Job Seeker

e) Student f) Homemaker g) Unemployed h) Disable i) Retired

❖ Question 4: How often do you use public transport?

a) Everyday b) Only on weekdays c) Occasionally d) I never use PT

e) Others: _____

❖ Question 5: What is your usual public transport mode when using public transport?

a) Train b) Bus c) Ferry d) Light rail

❖ Question 6: What purpose do you usually use public transport for?

_____

❖ Question 7: What time do you usually depart most often?

_____

❖ Question 8: What is your usual destination?

_____

❖ Question 9: From the scale of 1 to 10, please give your rating of the following infrastructure and service quality (with 1 being “Worst” and 10 being “Best”):

- Information offered: 1 2 3 4 5 6 7 8 9 10
- Cleanliness: 1 2 3 4 5 6 7 8 9 10
- Available assistance: 1 2 3 4 5 6 7 8 9 10
- Refreshment facility variety: 1 2 3 4 5 6 7 8 9 10
- Refreshment facility access: 1 2 3 4 5 6 7 8 9 10
- Toilet facility: 1 2 3 4 5 6 7 8 9 10
- Waiting benches at stops or stations: 1 2 3 4 5 6 7 8 9 10

❖ Question 10: From the scale of 1 to 10, please give your rating of your transport experience (with 1 being "Worst" and 10 being "Best"):

- Staff friendliness: 1 2 3 4 5 6 7 8 9 10
- Comfort during travel time: 1 2 3 4 5 6 7 8 9 10
- Temperature: 1 2 3 4 5 6 7 8 9 10
- On-board information: 1 2 3 4 5 6 7 8 9 10
- Vehicle cleanliness: 1 2 3 4 5 6 7 8 9 10
- Another customers' behaviour: 1 2 3 4 5 6 7 8 9 10

❖ Question 11: If you have experienced the public transport service during peak hour (normally 6-9 AM and 3-6 PM), please let us know about your experience.

_____

_____

_____

_____

❖ Question 12: Overall, from the scale of 1 to 5, please give us your rating of public transport service you usually use (with 1 being "Poor", 3 being "Average" and 5 being "Excellent"):

1 2 3 4 5

THANK YOU FOR YOUR RESPONSE!

HAVE A NICE DAY.