**EBA5001 DABP Day 3 Project Assignment**

Ang Khee Hwa Randi (A0198526N)

Chua Fu Lin Eugene (A0199046U)  
Koh Zhi Rong, Chester (A0125379X)  
Ong Kian Eng (A0052270U)

# Question 1: Student Competency Analysis

## Data Exploration and Results

We performed a simple data exploration to understand the data and asses its suitability to perform analysis. Based on our data exploration, the dataset does not have any missing values and could be used to perform analysis right away.



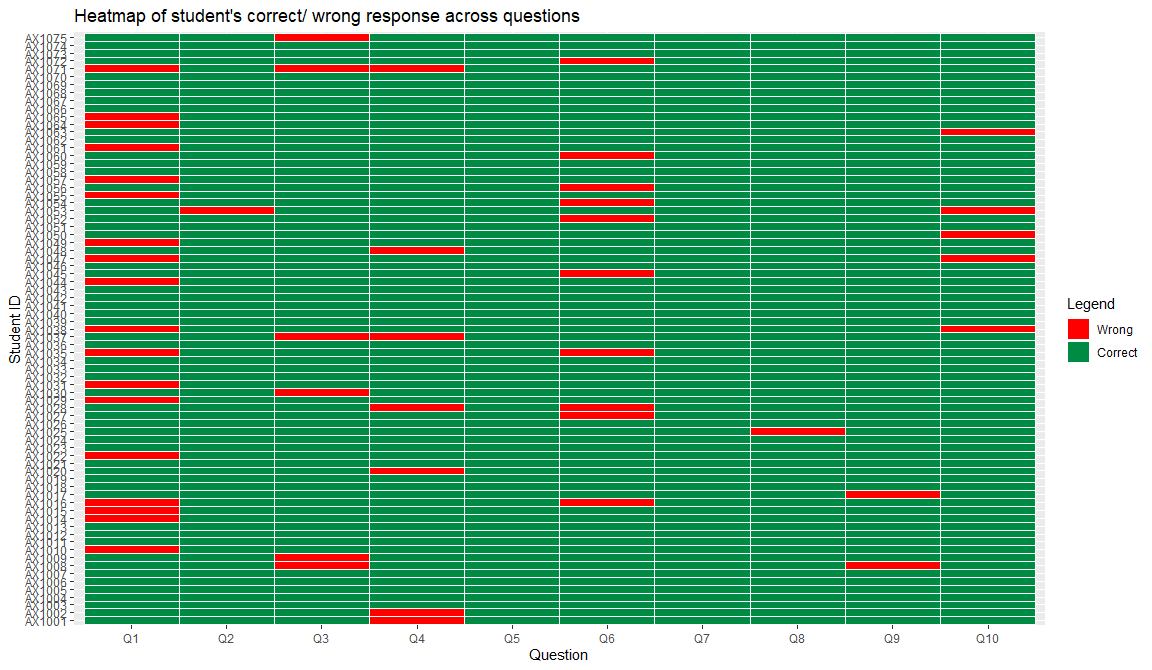
For exploratory purpose, we examined the different types of responses that were provided by students across all 10 questions. We stored the results of unique responses in a list for our reference so that we know what are the responses given by students for each question.

## [[1]]  
## [1] TRUE FALSE  
##   
## [[2]]  
## [1] "Numeric" "Integer"  
##   
## [[3]]  
## [1] TRUE FALSE  
##   
## [[4]]  
## [1] TRUE FALSE  
##   
## [[5]]  
## [1] "NA"  
##   
## [[6]]  
## [1] "Error" "Odd" "Even"   
##   
## [[7]]  
## [1] "Error"  
##   
## [[8]]  
## [1] "Error" "0"   
##   
## [[9]]  
## [1] "summary(df)" "str(df)"   
##   
## [[10]]  
## [1] "data.frame" "matrix" "Error" "list"

From the above output, Q10 got the broadest variety of wrong answers.

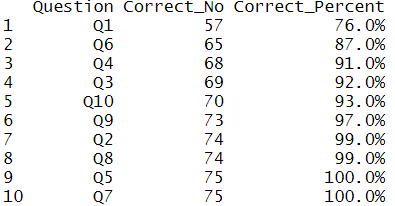
We can visualize how each student had fared for all 10 questions using a heat map. Besides providing a broad overview of the results for the cohort, the heat map also allowed us to, at one glance, identify:

1. questions that students often got it wrong
2. student with high proportion of wrong answers

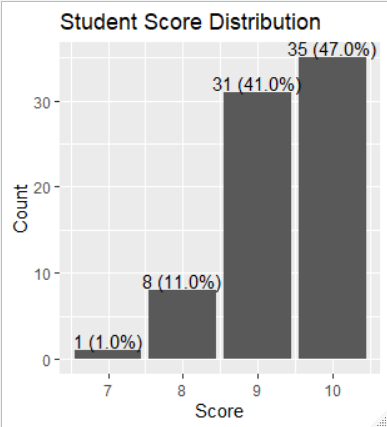


**Fig. 1: Score of each student for each question**

We also have created a table to show the number of correct responses and percentage of correct responses per question.



## Overall performance of students



We can see that majority of the students (88%) scored 9 and above. This group of students will be defined as high scoring candidates.

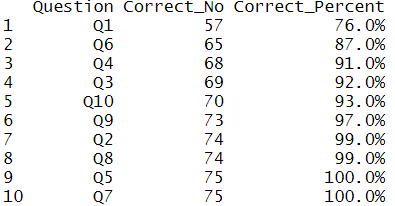
From educational perspective[[1]](#footnote-2), we can perform various item analysis for each question:

1. Difficulty index – Find out question students found difficult
2. Discrimination index – Find out which question different profiles of student have more problems with, and perhaps can re-use these questions for discussions for future batches or re-set the question in future to improve clarity of the question.
3. Distractor index – Find out the wrong options (when students are given options to choose from) that students often chose

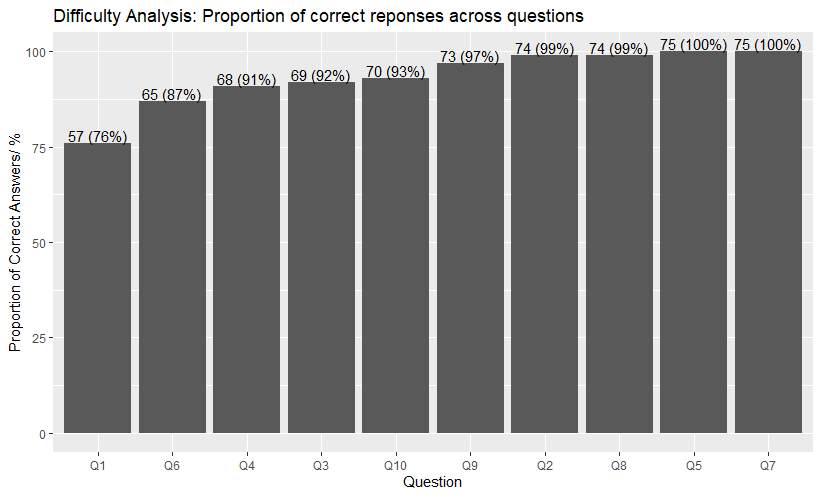
For the following analyses, we have chosen to illustrate questions that deserve more attention (i.e. those with higher frequency of wrong answers). The following chart provides a **visual** way to perform item analysis for each question **at one glance**.

## Difficulty Analysis

A summary table of all the questions and proportion of students who got the questions correct will allow us to see how the cohort fared.

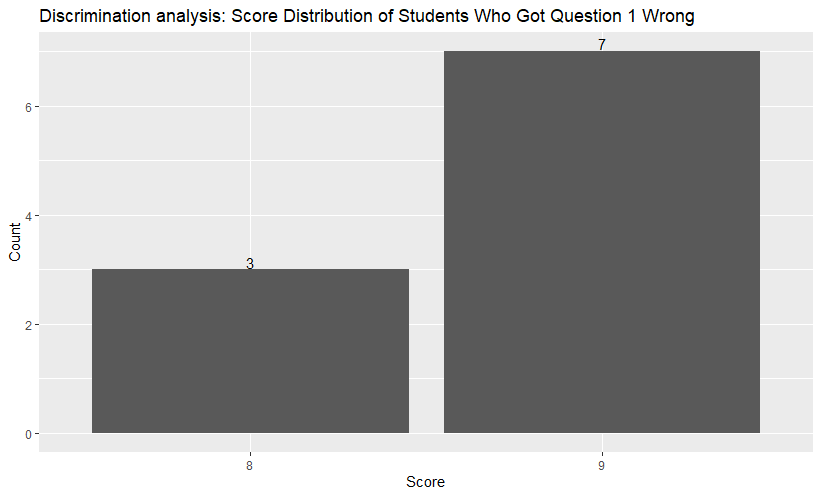


However, having a bar plot as compared to table will allow us to instantly and precisely identify the questions that deserve greatest attention. This allows us to determine the difficulty level of each question by percentage of students who got the correct answer. Students generally perform well across all questions except for Question 1 and 6.

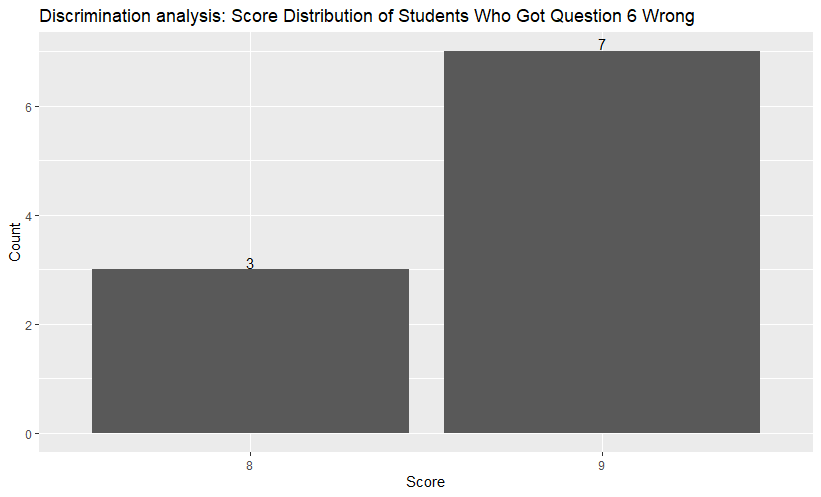


## Discriminator Analysis

From the bar chart, we can perform a visual distractor analysis to see the frequency of various responses given by high-scoring candidates (i.e. total 9 points) and low-scoring candidates (i.e. 7 and 8 points). This allows us to, at one glance, have a sense of the difficulty of the question.



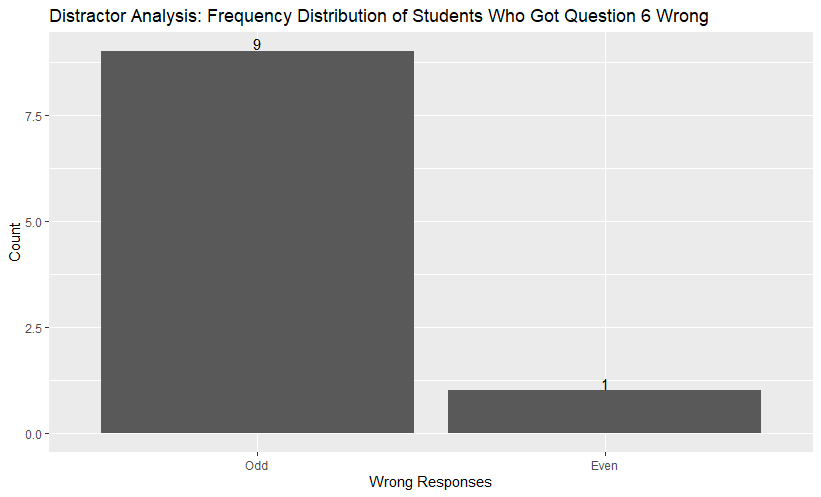
From the distribution, we can see that Question 1 is a good discriminator. Not only did the low-scoring candidates (e.g. total score of 7) got this question wrong, high-scoring candidates (total score of 9) also got it wrong. This meant that the stronger students also had difficulty answering this question. This could be due to students’ conceptual error (e.g. misconception or alternative conception), or problem with the question (e.g. clarity of question) that resulted in different interpretations and led to various responses given.



Similarly, numerous high-scoring candidates (total score of 9) also got Question 6 wrong.

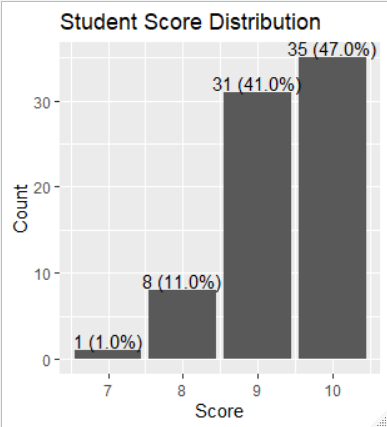
## Distractor Analysis

We can visualize the frequency of various wrong responses that students gave so that we can address the common mistakes (in terms of higher frequency of wrong response) made by students.



From the bar chart, we can see that most students chose the wrong response “Odd” for question 6, hence the course manager can then address why the response “Odd” is the wrong answer during review lesson.

## Overall Comments



In conclusion, most students have generally done well for the survey as most students (88%) score 9 and above. From the charts above and Fig. 1, we can see that questions 1 and 6 are difficult questions. Hence, the course manager should spend more time going through the concepts in questions 1 and 6. Additionally, we could also see that there is one student who is lagging behind the class, perhaps more attention and guidance could be given.

# Question 2: Ride Sharing Analysis

## Assumptions made

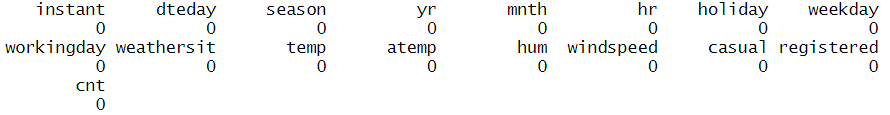
For the ride sharing analysis, we would be making the following assumptions:

1. We would be taking on the role of an operations manager for the **bike sharing** company to monitor the health of the firm’s operations and key variables that may influence bike ridership.
2. The firm works on a pay per use model of its bikes to generate revenue, hence there is a need to maximize its bicycle utilization for a higher return on investment.
3. There are distinct seasons (spring, summer, autumn and winter).

Our analysis will focus on understanding ridership patterns and any possible relationship between ridership and other factors. We are at the first stage of our analysis where it will be mostly descriptive. Any possible relationships identified will be subjected to statistical testing after discussion with the management team.

## Data Cleaning and Preparation

The dataset was checked for missing values. There is no missing value.



As the data is in hourly format, data was aggregated at daily level.

## Business Question 1: What is the general trend of bicycle ridership over the span of a year?

Finding out such a trend would allow us to identify the performance of the firm’s operations, as ridership numbers would be a proxy for the potential revenue generated for the firm. This may also shed light on the possible impacts and effectiveness of changing business conditions and business strategies on ridership respectively. The nuances of any pattern uncovered will allow for better resource planning. Furthermore, we can also explore whether seasons (spring, summer, autumn and winter) within the year may affect ridership.

### Findings

As there is too much volatility daily, the aggregated monthly ridership will be used for analysis instead. Based on our findings in Figure 1, bike ridership is at the highest from June to August, and lowest from January to March. There is the steep rise in ridership in the first quarter (season 1, from January to March) to the second quarter (season 2, from March to June). Ridership peaks in June. This is followed by a gradual decline in the subsequent quarters. Given the state of decline not fully matching up with January’s baseline ridership figures, we suspect that there was a major event that had occurred which boosted ridership in in quarters 1 and 2. However, there is a lack of information provided in the dataset to determine cause of the sharp increase – this could be attributed to a change in marketing strategy or rapid expansion by the firm. Also, as the dataset is limited to just 1 year, we are unable to see any seasonality (i.e. repeatability / consistency of trends). If the decline in ridership in the later 2 quarters is not due to seasonal effects, this can be a worrying sign for the health of business, as ridership (sales) is declining which decreases our revenue.

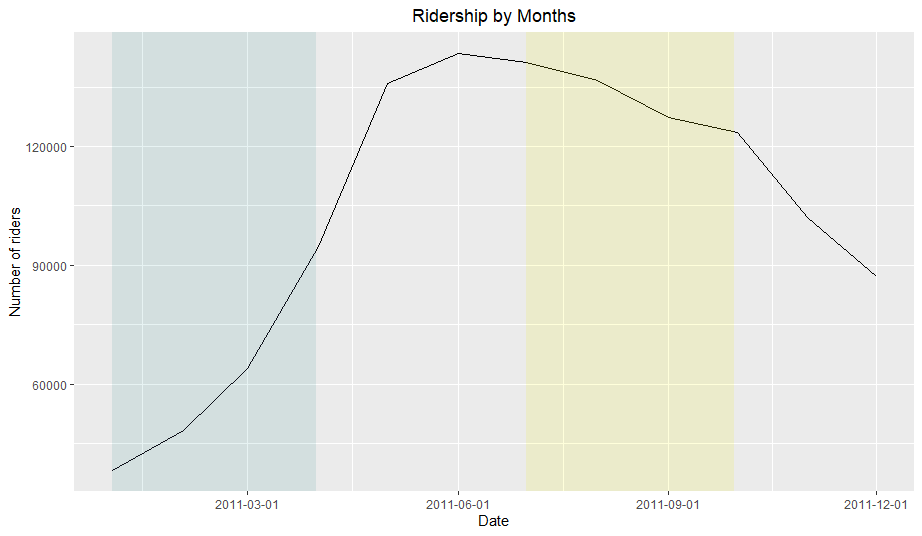


Figure 1: Bike ridership in 2011. A noticeable increase in observed between the first and second quarter of the year, peaking during the summer period in June, before gradually declining for the rest of the year.

**Additional Notes on Figure 1:**

Different months have slightly different number of days (each month has between 28 to 31 days), which may result in a variation in total ridership between months. However, in this case the ridership trend is pronounced enough, such that its effect is relatively negligible.

## Business Question 2: What are the key variables that may affect bicycle ridership?

Identifying key variables (factors) that may affect bicycle ridership would allow us to maximize or minimize their impact on our ridership, so that we can better pre-empt ridership, leverage on insights and opportunities, and allocate resources optimally (e.g. marketing or publicity at the right time / opportunity). We can also fine-tune business operations and strategies accordingly.

### **Findings**

We conducted a correlation analysis for various variables to understand how each variable is correlated with the number of registered and casual riders. As the categorical variables (e.g. season, weather situation) are ordinal variables, we have included it in our analysis to glimpse into any potential correlation between them and ridership (although this may not be the most accurate method).

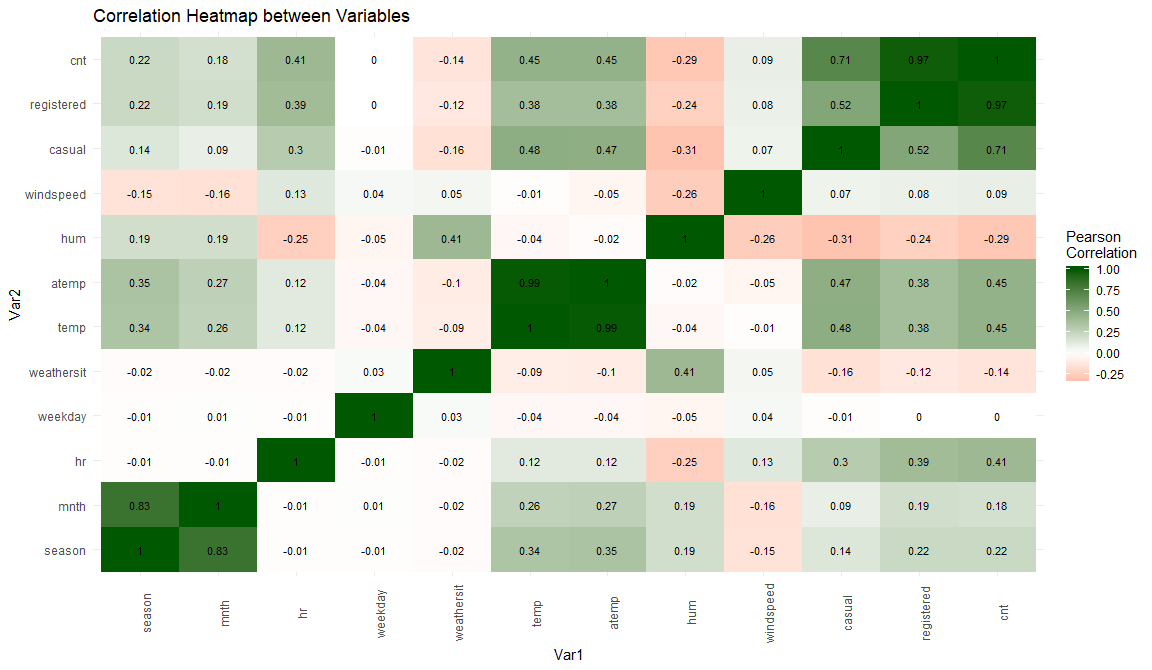


Figure 2: Correlation Map. Temperature is observed to have the strongest correlation with bike ridership as compared to other variables.

Based on our findings in Figure 2 and Figure 3, we found that (in order of decreasing strength of correlation):

1. Temperature (including normalized feeling temperature) has a positive correlation to total bike ridership (+0.45).
2. Time of the day has a positive correlation to total bike ridership (+0.41)
3. Humidity has a weak negative correlation to total bike ridership (–0.29)
4. Season has a weak positive correlation to total bike ridership (+0.22)
5. Weather situation has some weak negative correlation to total bike ridership (–0.14)
6. Windspeed has a weak positive correlation to total bike ridership (+0.09)

Among the various numeric variables, temperature (including normalized feeling temperature) has the strongest correlation to total ridership. Hence, from operations and marketing points of view, we will focus on investigating how normalized feeling temperature affects bike ridership. The normalized feeling temperature may be a better variable to use as compared to actual temperature because the feelings of the riders (i.e. what the temperature feels like because from the rider’s standpoint, comfort level can be related to a combination of factors – temperature, humidity and windspeed[[2]](#footnote-3)) will likely to affect the riders more than the actual temperature.

Nonetheless, we need to be particularly wary of including both these variables in any model building as there may be multicollinearity issues.

## Business Question 3: How does the bicycle ridership change within a 24-hour period for different days of the week – between working weekdays and non-working days (weekends and holidays)? Also, how does the usage patterns differ between registered and casual users?

We hypothesize that weekday ridership would see a bimodal peak – one in the morning and another in late afternoon to early evening periods of weekdays. This will coincide with the timings of daily weekday peak hour commutes for working adults and students. We are likely to see more registered users.

On the other hand, a holiday or weekend may not have the bimodal peak since most commuters are not commuting to school or work. We are likely to see more casual users too.

Hence, from operations and marketing points of view, we can optimize operations accordingly, such as planning:

1. Off-peak times to service bicycles
2. Load distribution of bicycles by shifting them to key locations[[3]](#footnote-4) with higher demand to ensure adequate bicycles are available for use during peak periods
3. Promoting usage during off-peak hours to achieve greater revenue

The hourly plot across different days will allow us to see if there is a repeated pattern.

Registered and casual users may have different usage patterns. As most of the customers are registered customers, it would vital for us to understand their usage patterns so that we are able to serve them better and achieve higher rate of customer retention for long-term revenue.

Understanding casual users’ behaviors will allow us to better plan our resources and to maximize utilization of bicycles during off-peak periods (i.e. load distribution) or use it for maintenance. More importantly, it will be essential for us to better understand when (e.g. based on certain type of weather condition) are they likely to use our bicycles. This will be useful in getting them convert to become our registered users.

### **Findings**

The table below concurs with our hypothesis that there are more registered riders on working weekdays and more casual riders on weekends.

Table 1: Mean ridership on weekdays vs weekends. Casual riders are observed to prefer riding on weekends, whilst registered riders are observed to ride more on weekdays. Holidays have been excluded from the analysis to better normalize the values.

|  |  |  |  |
| --- | --- | --- | --- |
| Type of day | Total Riders (Mean) | Casual Riders (Mean) | Registered Riders (Mean) |
| Working weekdays | 171253 | 23671 | 147582 |
| Weekends | 178408 | 59575 | 118834 |

From the table below, we can conclude that Fridays, followed by Mondays and Tuesdays are the most popular days for total bike ridership. For Casual riders, the most popular days are weekends whilst Registered riders ride more on weekdays (Tuesdays, Fridays, Thursdays), as seen in Table 2.

Table 2: Most popular day of the week by ridership. Casual riders are observed to prefer riding on weekends, whilst registered riders are observed to ride more on weekdays.

|  |  |  |
| --- | --- | --- |
| Total Riders | Casual Riders | Registered Riders |
| weekday sum  *<int>* *<int>*  1 5 182006  2 2 180338  3 1 180221  4 6 179743  5 0 177074  6 4 174552  7 3 169169 | weekday sum  *<int>* *<int>*  1 0 59603  2 6 59547  3 1 31560  4 5 29453  5 2 23989  6 4 22440  7 3 20660 | weekday sum  *<int>* *<int>*  1 2 156349  2 5 152553  3 4 152112  4 1 148661  5 3 148509  6 6 120196  7 0 117471 |

The observation above concurs with our initial hypothesis. Registered riders may be riding more on weekdays, primarily to school or work in the morning and back home in the late afternoon or evening, as seen in the bimodal ridership pattern in the morning and late afternoon on weekdays as seen in Figure 4. Their heavy use of bike rentals is likely to be a reason for them to join our subscription as registered members. Casual users, on the other hand, casual users rents bike for ad-hoc commutes and leisure as seen from the single peak observed for each day and spikes in ridership on weekends in Figure 3.

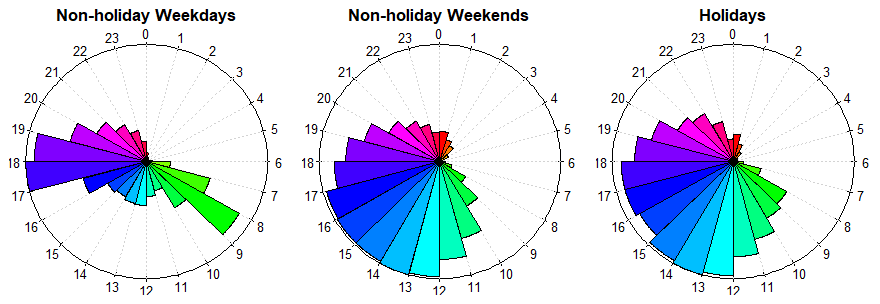


Figure 3: Bicycle ridership pattern for all riders on workdays (Mondays to Fridays), rest days (Sat and Sunday) and holidays. Holidays and rest days are observed to have spread out ridership patterns through the daylight hours, whilst workdays have ridership concentrated in the morning and afternoon peak periods.

*(Note: Time shown is in 24 hours format. 0 refers to midnight)*

To better understand the hourly pattern of ridership for each day of the week, line graphs for each day of the week were plotted.

From the graphs below (Figure 4 and Figure 5), during weekdays, there is bimodal peak ridership pattern in the morning and late afternoon for registered users. On the other hand for weekends, there is a single peak ridership pattern for casual users as seen in Figure 5. This concurs with our hypothesis earlier on.

Furthermore, majority of the bike riding during daylight hours, and fewer occur at night which could possibly due to greater safety concerns due to lower visibility at night.

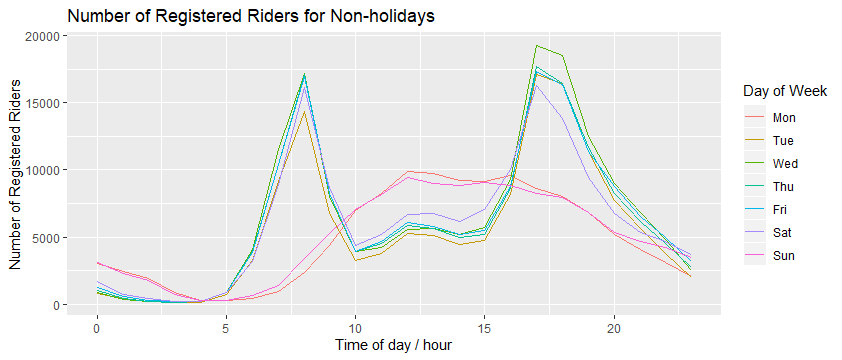


Figure 4: Bimodal peaks ridership pattern of registered riders in a 24-hour timespan for each day. Registered riders tend to ride predominantly on the morning and late-afternoon peak periods on weekdays.

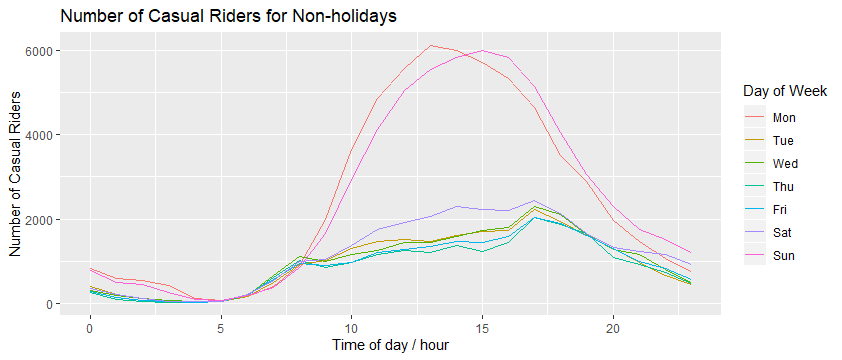


Figure 5: Single peak ridership pattern of casual riders in a 24-hour timespan for each day. Unlike registered riders, casual riders ride predominantly on weekends.

## Business Question 4: How does temperature impact ridership?

From our earlier analysis, temperature has some correlation to ridership. For example, ridership is likely to be higher in warmer temperatures. Seasons (spring, summer, autumn and winter) which affect temperature will likely impact ridership. For example, warmer seasons in spring and autumn would make more pleasant rides and hence see higher ridership, as compared to cold autumn and winter seasons. Depending on the season and temperature, we may strategize and focus our business operations in areas with warmer temperatures; or use the lull periods in ridership to perform more bicycle maintenance or rent the bikes out for other events to gain more revenue.

In addition, with the increasing temperature due to climate change, we could also pre-empt the impact of climate change on ridership.

### **Findings**

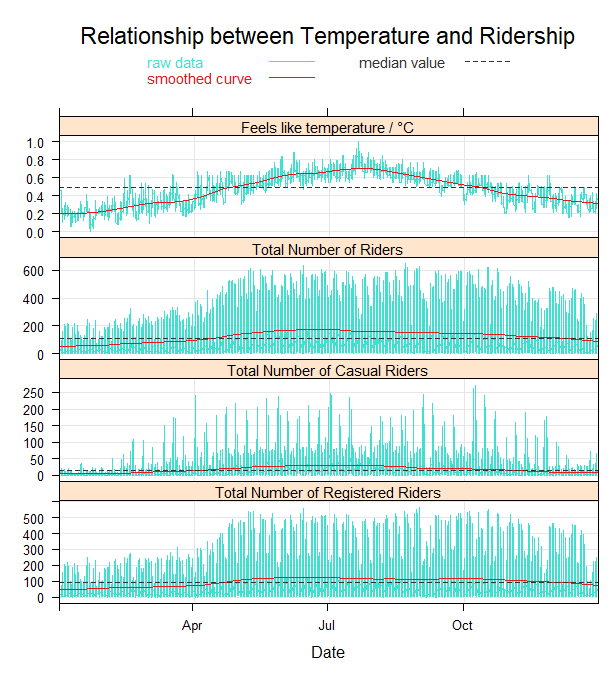


Figure 6: Bicycle ridership and temperature trends across 2011. There appears to be a positive correlation of temperature and bicycle ridership over the span of the year, though bike ridership it not very sensitive to temperature changes.

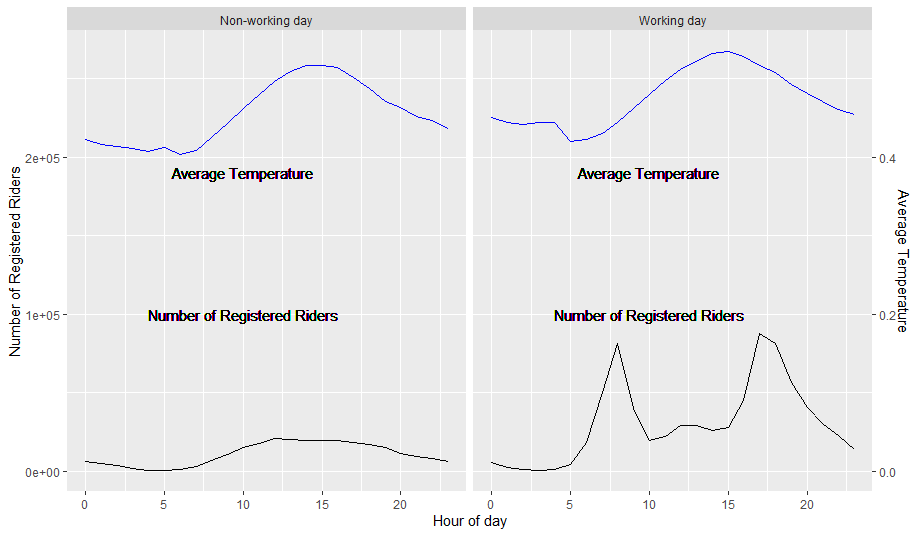


Figure 7: Bicycle ridership of registered riders vs average temperature for hour of day. The pattern of hourly temperature of each day does not coincide with that of registered users’ ridership.

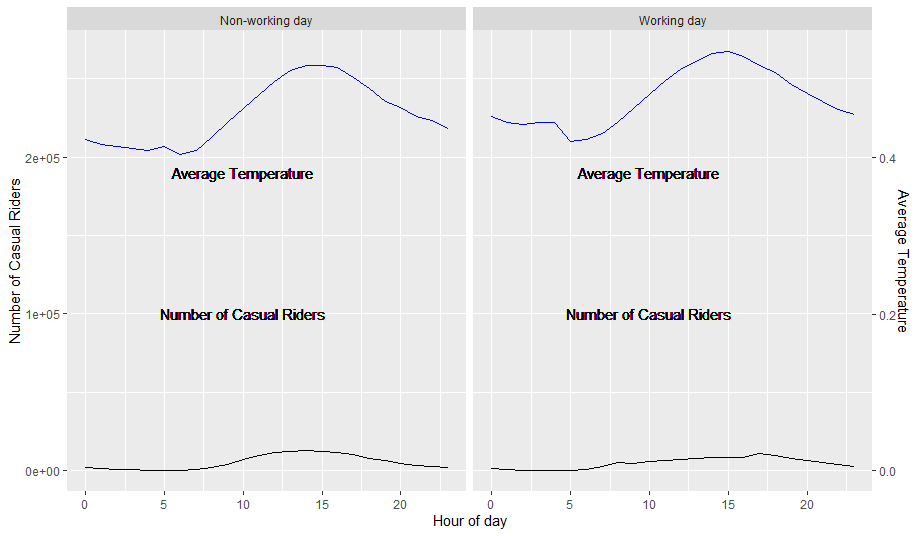


Figure 8: Bicycle ridership of casual riders vs average temperature for hour of day. The pattern of hourly temperature of each day does not coincide with that of casual users’ ridership.

Based on Figure 6, we can tell that there are more riders during warmer temperatures, especially across the seasons. Colder seasons (autumn and winter) have comparatively fewer riders than in warmer seasons (spring and summer).

However, there is a lack of clear correlation between ridership and changes in average hourly temperature in Figure 7 and Figure 8.

## Business Question 5: How do different variables under different weather situations impact ridership?

Given that we have already observed an inverse relationship between weather situation and ridership in the earlier correlation heatmap, we would like to explore it in further detail to determine the optimal weather conditions for ridership. Based on this inverse relationship, ridership will be highest on clear days, and progressively decreases as weather worsens (e.g. misty days or days with precipitation – rain or snow), perhaps due to increased safety concerns and discomfort from being wet or cold. Hence, by gaining more insights on this, we can take actions such as relocating bikes to places with better weather, in order to maximize ridership for the day.

The area chart was used instead of line graph so that we can see any distinctive patterns much more clearly.

### **Findings**

Based on the graphs and table below, we found that the weather situation has a stronger correlation to ridership, as compared to other variables. Days with better weather (i.e. Clear in weather situation 1 and Misty in weather situation 2 weather) would result in higher ridership as compared to days with poor weather such as heavier precipitation (weather situations 3 and 4). Unsurprisingly, there are almost no rides on days with heavy precipitation (weather situation 4). However, it should be noted that registered riders are less elastic to changes in weather, possibly due to their need to commute to work or school, whereas casual riders tend to ride for more personal reasons as seen from their more spread out ridership pattern that spreads out across the day and spikes on weekends.

Using the sum values will introduce bias to the analysis in favour of better weather situations (1 and 2) as there are more instances of good weather than poor weather (3 and 4). As such, sum values of ridership were not chosen. Hence, we decided to look at average or median. As we do not know if the distribution of riders in different conditions is normally distributed, we decided to choose median.

Table 3: Median ridership number for each weather situation. Overwhelmingly, ridership is highest on days with good weather (weather situation 1) and decreases as weather deteriorates progressively from weather situation 2 to 4.

|  |  |  |  |
| --- | --- | --- | --- |
| Weather Situation | Median Total Riders | Median Casual Riders | Median Registered Riders |
| 1 – Clear | 17 | 122 | 98 |
| 2 – Misty | 13 | 106 | 87 |
| 3 – Light Rain / Snow | 4 | 50 | 44 |
| 4 – Heavy Rain / Snow | 1 | 36 | 35 |

From the following graphs which compared median ridership against other weather variables such as temperature, humidity and windspeed by weather situation, we were unable to observe a clear and distinct relationship of their impact on ridership that will help us formulate strategic action plans to leverage on certain weather conditions to promote ridership. However, we notice that there is completely no ridership during heavy rain and snow conditions. Using this information and weather forecast, we could plan to reallocate resources to focus on bicycle maintenance rather than redeploying bicycles to areas with high usage (bicycle re-distribution). This will save manpower cost from bicycle re-distribution. It will also optimize maintenance and minimize disruption to our customers, as maintenance is carried in the least disruptive period. Maintenance cost could also be potentially reduced if we retrieve some of our bikes so that they would not be left out in the open and be damaged in harsh weather conditions. Likewise, during lull period of light rain or snow, misty conditions, we could also carry out some maintenance work.

**Temperature**

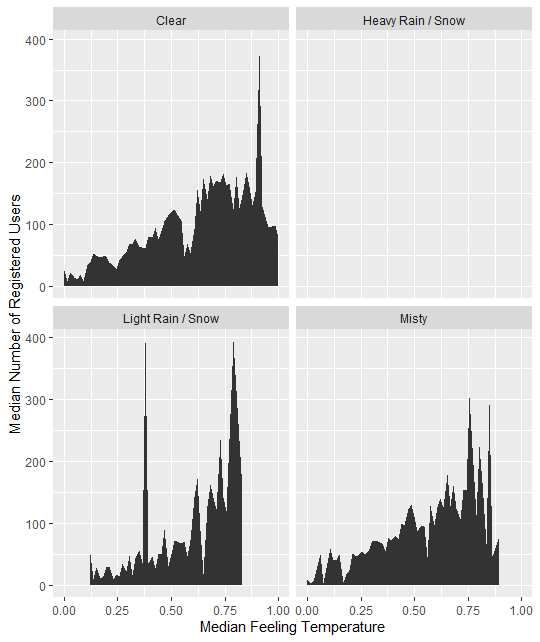


Figure 9: There is no clear relationship observed between median temperature and median registered ridership for each weather situation. Ridership seems to increase slightly with temperature.

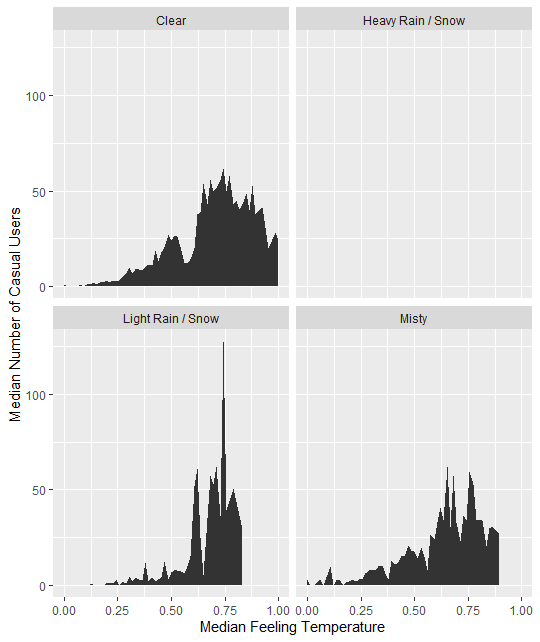


Figure 10: There is no clear relationship observed between median temperature and median casual ridership for each weather situation. Ridership seems to increase slightly with temperature.

**Windspeed**

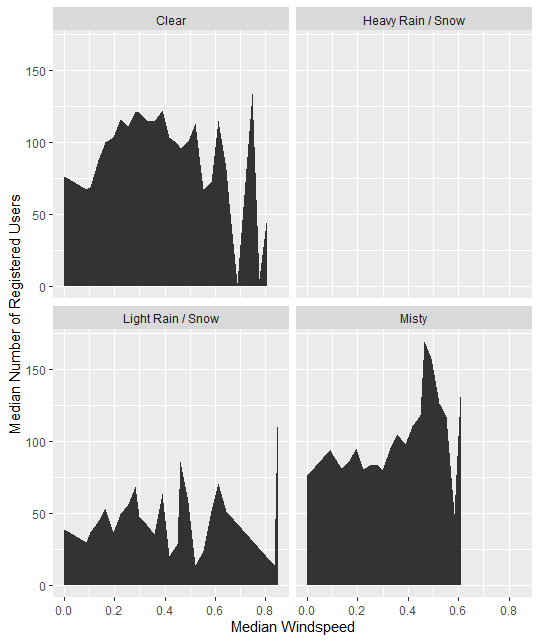


Figure 11: There is no clear relationship observed between median windspeed and median registered ridership for each weather situation.

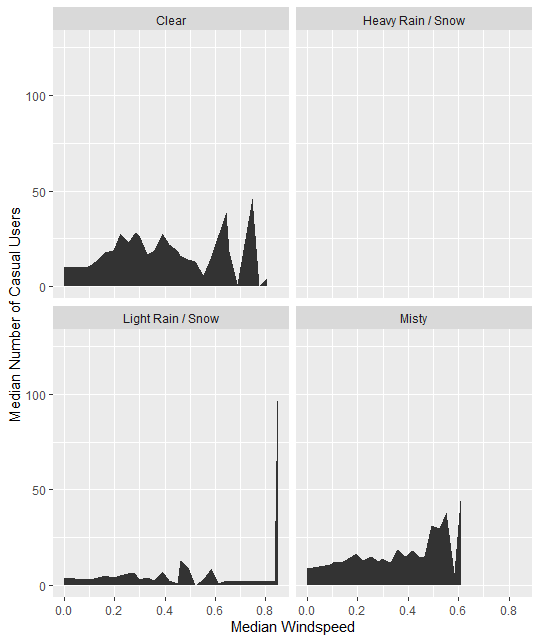


Figure 12: There is no clear relationship observed between median windspeed and median casual ridership for each weather situation.

**Humidity**

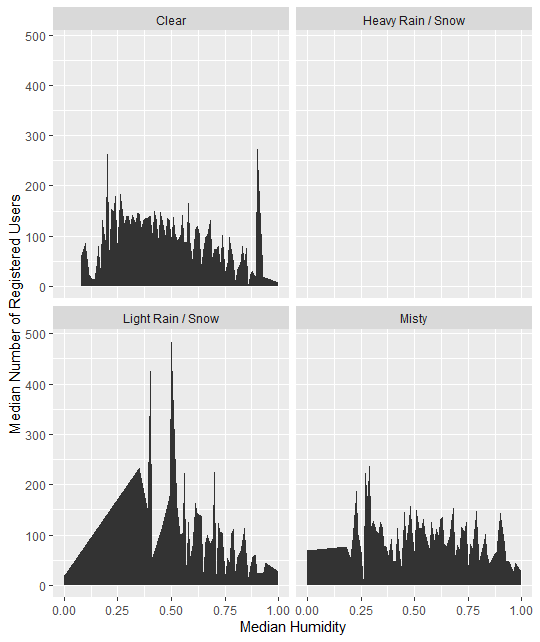


Figure 13: There is no clear relationship observed between median humidity and median registered ridership for each weather situation. However there seems to be a significant portion of riders who are willing to ride in humidity levels of around 0.25 to 0.50 in weather situation 3. These are likely riders who are willing to brave through light rain / precipitation (which has higher humidity levels) for their commute.

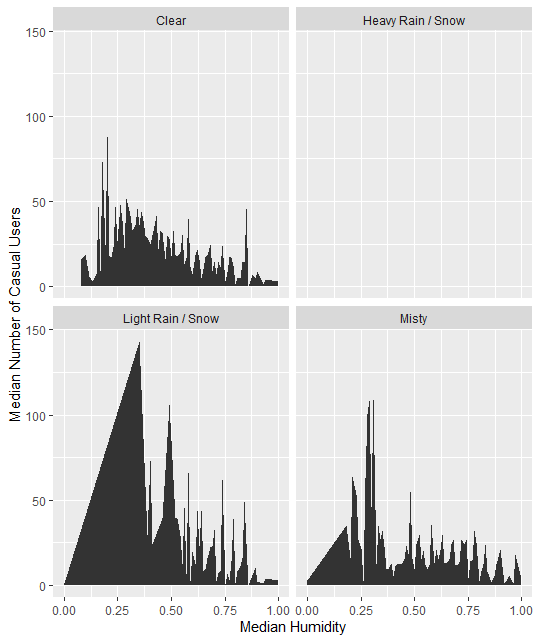


Figure 14: There is no clear relationship observed between median humidity and median casual ridership for each weather situation. However there seems to be a significant portion of riders who are willing to ride in humidity levels of around 0.25 to 0.50 in weather situation 3. These are likely riders who are willing to brave through light rain / precipitation (which has higher humidity levels) for their commute.

We further classified the different seasons and weather situation to see if there is any pattern. However, from the following graphs, we were still unable to observe a clear and distinct relationship of their impact on ridership. We could only observe the differences in the number of riders across seasons.

**Temperature**

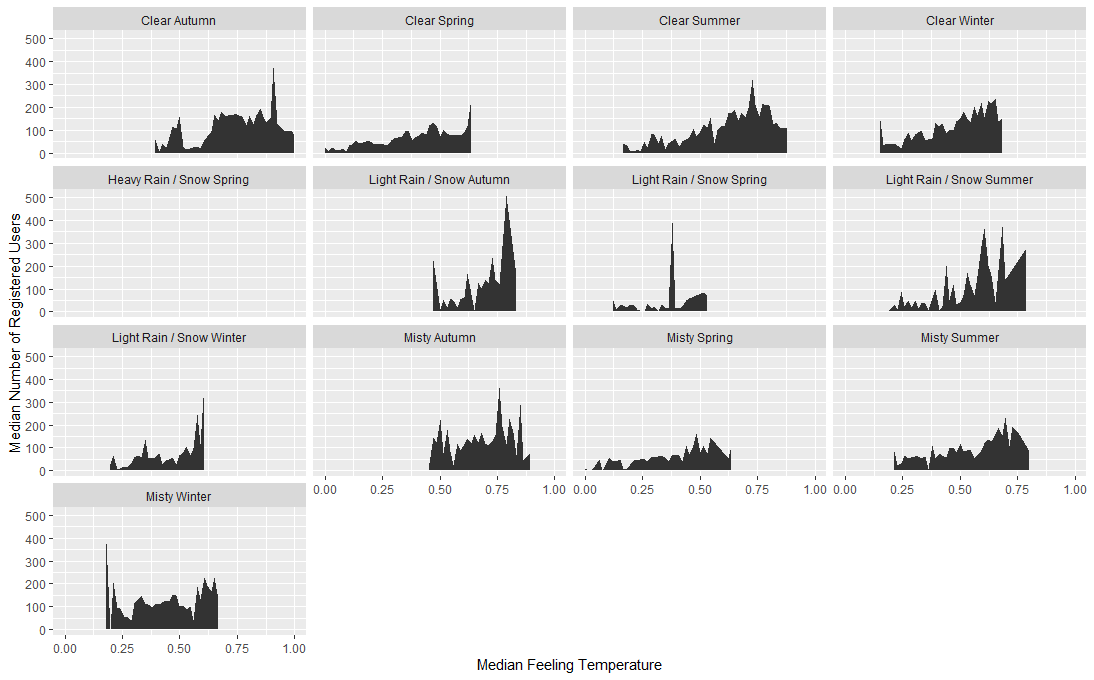


Figure 15: There is no clear relationship observed between median temperature and median registered ridership for each weather situation in each season.

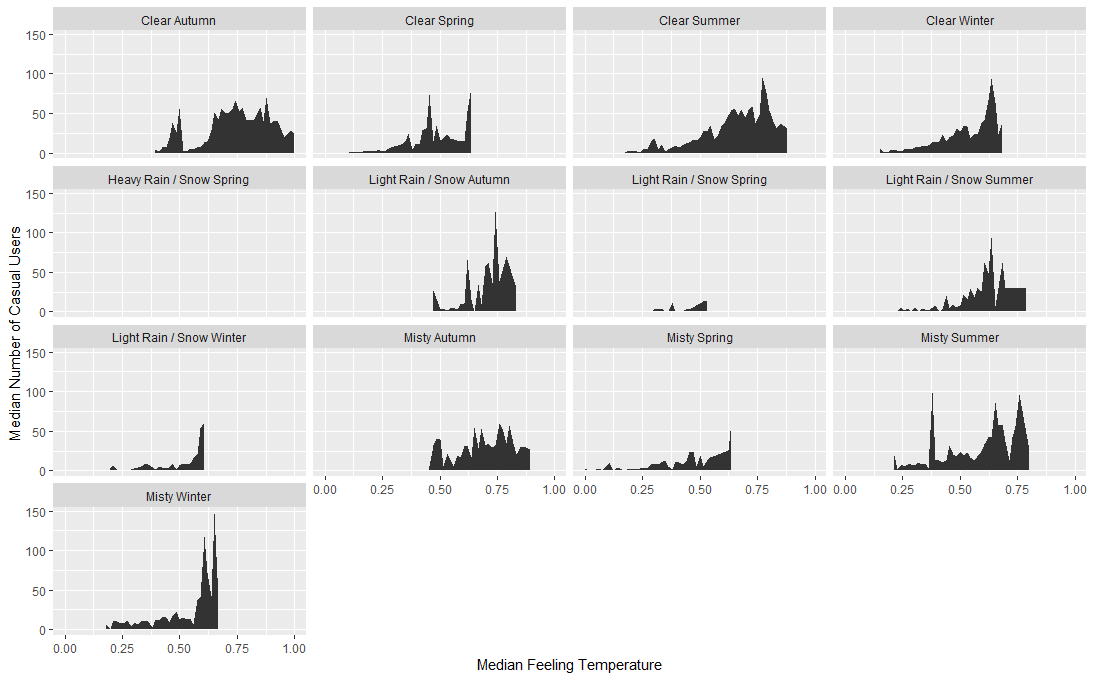


Figure 16: There is no clear relationship observed between median temperature and median casual ridership for each weather situation in each season.

**Windspeed**

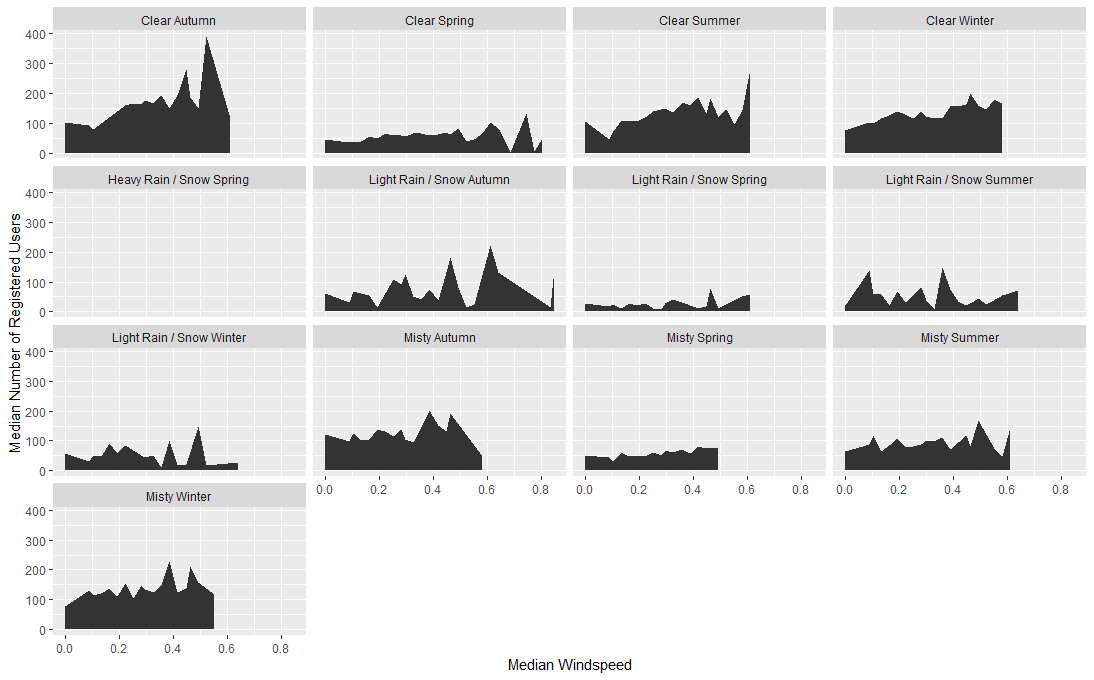


Figure 17: There is no clear relationship observed between median windspeed and median registered ridership for each weather situation in each season.

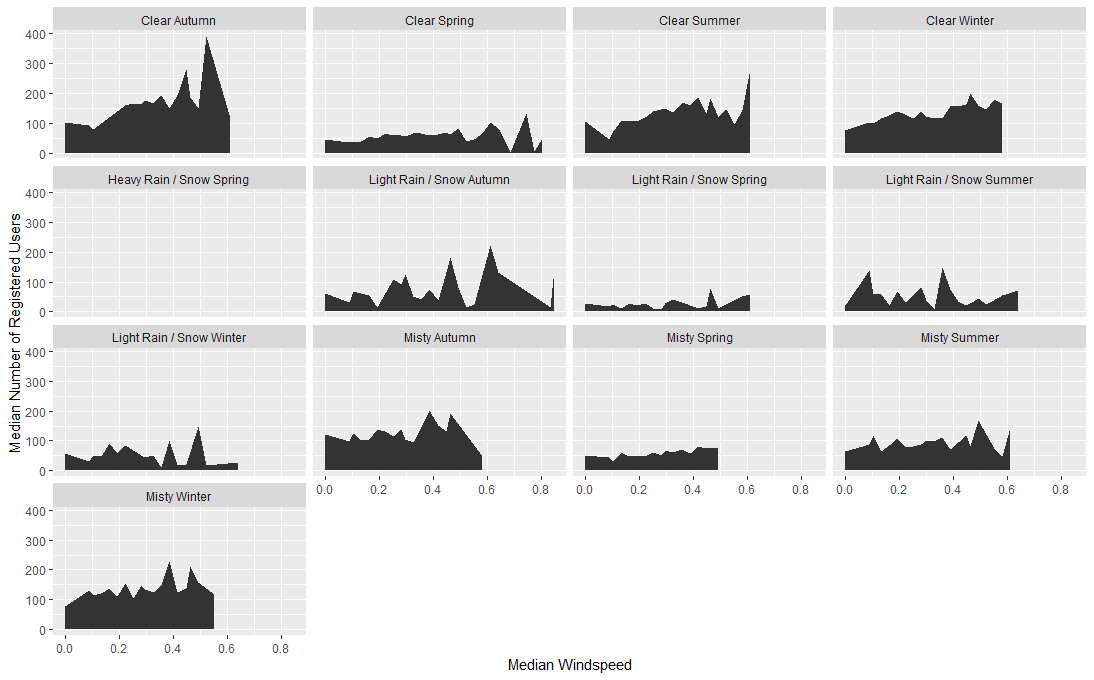


Figure 18: There is no clear relationship observed between median windspeed and median registered ridership for each weather situation in each season.

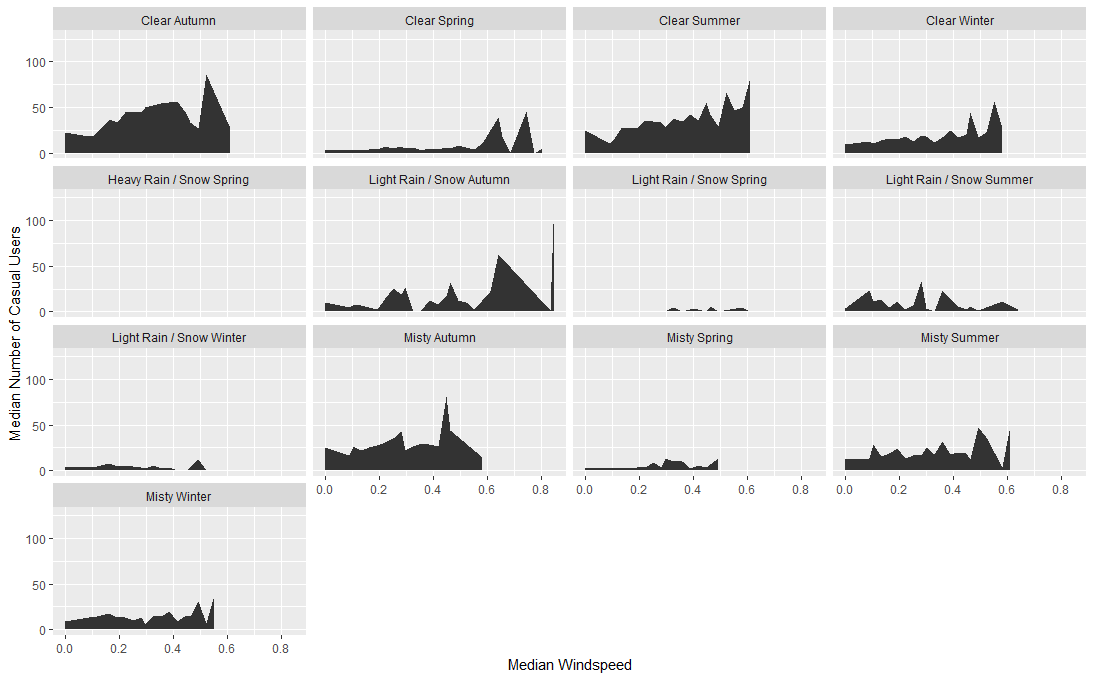


Figure 19: There is no clear relationship observed between median wind speed and median casual ridership for each weather situation in each season.

**Humidity**

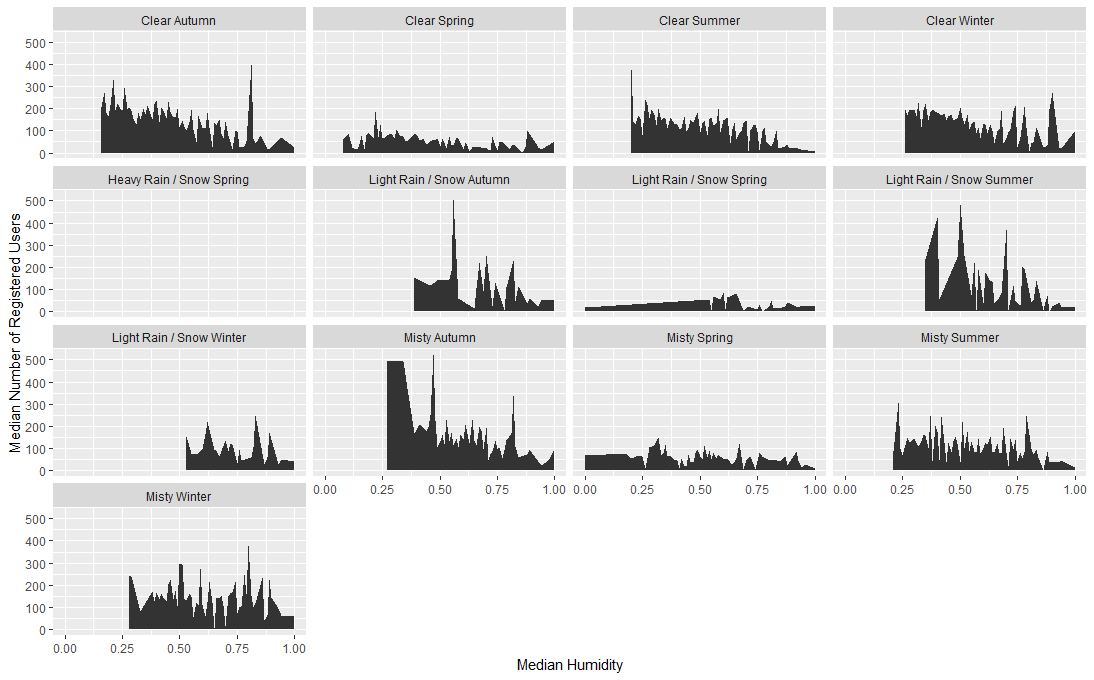


Figure 20: There is no clear relationship observed between median humidity and median registered ridership for each weather situation in each season.

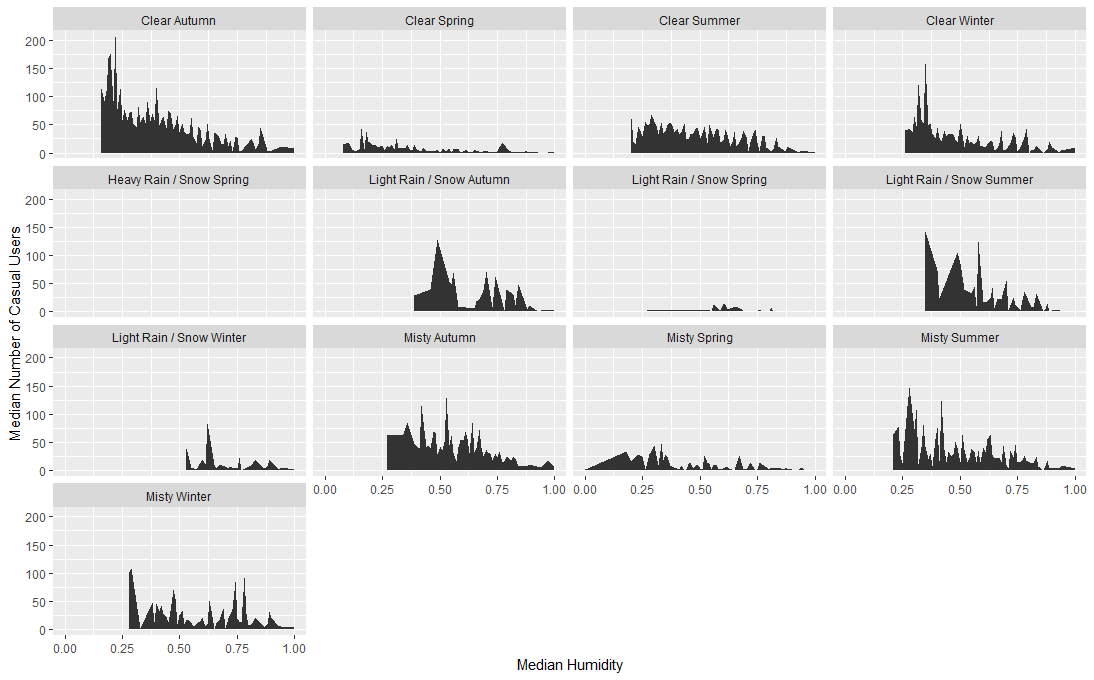


Figure 20: There is no clear relationship observed between median humidity and median casual ridership for each weather situation in each season.

1. Reference / Further explanation on various indices:

   <https://www.education.gov.za/Portals/0/Documents/Publications/NEEDU%20POLICY%20BRIEF%20SERIES/Policy%20Brief%207A.pdf?ver=2018-07-02-103215-443> [↑](#footnote-ref-2)
2. Example: In cold countries, even if the temperature is warmer (e.g. 10°C), a high humidity and strong windspeed (wind chill factor) may make it uncomfortable for riders to ride. [↑](#footnote-ref-3)
3. Location data needs to be factored in for more detailed analysis. However, location data is not provided in this dataset, limiting further analysis. [↑](#footnote-ref-4)