Bike Sharing Services: A statistical analysis report

Mark Movh

ABSTRACT

The following statistical report investigates a US-based dataset revolving around bike-sharing services spanning from 2011-2012. From this dataset, three questions arose in regards to environmental factors, season and weather, and lastly in regards to certain groups. The purpose of this research was to gather insightful information to better understand bike-service demands and patterns, which can be used for future prediction and better implementation of said system. The report applies several mathematical techniques, ranging from correlation to 2-sample tests, and data visualisations, including: heatmaps, scatterplots, line-charts, bar-charts, boxplots, pie-charts and tables. Findings from these visualisations and techniques found some correlation between temperature and cyclist count, seasons of summer and autumn being most popular, clearer days having more cyclists on average, with autumn having the clearest days. Furthermore, regulars were more active than casuals at 8-9am and 16-17am, while casuals saw an increase towards the afternoon.

1 INTRODUCTION

Biking had always been an alternative to transport, paving way for more efficient and health methods to get from one point to another. In terms of efficiency, climate change has been becoming increasingly prominent in day to day lives. Through the use of bicycles, it is possible to reduce emissions otherwise released by vehicles. Decreasing a person's carbon footprint left on the environment. Recent studies have shown that increased number of cycling results in a decrease in vehicle-related CO_2 emissions. Saving 1 trip a day, for at least 200 days a year reduces half a tonne of CO_2 per year. [2] When biking, the extra benefit is the additional exercise, allowing a person to be more active. The additional health benefits as a result can be summarised as decreases in various types of morbidity, including obesity, cardiovascular and cancer through cardio-respiratory fitness. [7] Bringing forth the premise that biking presents both physical and mental benefits.

However, not everyone has access to a bicycle at disposal for their personal use. For either monetary reasons or sake of inconvenience (Perhaps living too far away), not everyone has access to a bike. In 2015, over 800 large municipalities installed a bicycle sharing system, which is fairly large compared to 2004, where only 13 large municipalities had systems for such public access to bikes. Within almost 15 years (2000-2014), cyclist amounts rose by 62% as people bike to work in the US alone. [3] With this public access, it gives more people option to use a bike and become sustainable, while additionally generating some revenue through an automated system. Such systems allow for rental of bikes either for casual use or through some membership implementation. Due to the visible trend, and the increasing demand for bikes, this statistical analysis report investigates and analyses patterns between various environmental factors, groups of bikers and time intervals. In order to draw conclusions, and provide further insight into the use of bike services, this report looks into potential correlations, visualisation of data and generating discussions on findings.

2 DATASET DESCRIPTION

The statistical analysis is based upon data extracted from the "bike sharing dataset" [4] which overlaps the years 2011 and 2012. The log of information was received from Capital Bikeshare system, located in the United States, in Washington D.C. The dataset is mainly present in the 2013 "Progress in Artificial Intelligence" Journal. The dataset consists of time intervals (seasons, months, days, hours), environmental factors (temperature, humidity, wind-speed) and count for groups of bikers (casual, registered). Several questions arise from this dataset, which can be used to extract valuable information. The report primarily focuses on investigating several the data fields simultaneously; hence the following questions were generated:

(1) To what magnitude can correlations be seen existing between various environmental factors and cyclist count?
(2) To what extent does season and weather-type display some effect on bike services demands?
(3) How do business/school hours imply a difference between the casual and registered group of cyclists?

These questions weren't chosen at random. The questions were created in a way that could explore many of the data fields, and look into them. Such data could be gone over, and in turn hopefully present information that may be deemed useful. Either for reasons of practicality or further investigation/research.

Question 1 was chosen to investigate the environmental variables found in the dataset. Temperature, humidity and wind-speed were chosen specifically to see if any correlations exist between not only each other (to gain insight towards weather data) but also potential correlations to number of cyclists during those times. The importance of this question can help visualise how certain atmospheric conditions impact number of cyclists. This can then be interpreted and analysed to gain a better understanding of a biker's mentality, of when they are most active or when will they most likely to borrow a bike. Predictions can be created based on the information, to further improve the system.

Question 2 is of interest because seasons with better weather conditions and more ideal temperatures do tend to lead to more people being active, having more time off and taking vacations which can all lead to higher demand requirements on bikes. It is important to keep track of the numbers, and how they are seen increasing so it can be better predicted how many bikes and how high their availability should be during the busier times with better weather conditions.

Question 3 was deemed intriguing to explore because it can be important to see how the number of casual cyclists is affected during the busiest hours of the day, since it can be assumed that less people have time. However, the number of registered cyclists may have an opposite effect because they use bikes to get to school/work. Hence it can lead to answering questions of not only when the bikes should be set up, hour wise, but also which group has the higher demand. Furthermore, perhaps with more data it could be possible to locate hot spots.

3 DATA ANALYSIS

3.1 Question 1

Hypothesis: There will be little correlation among the actual environmental factors, at least between wind and other factors because unless there are dangerously high speeds, it shouldn't impact number of bikers for example. However, when temperature is ideal then the number of bikers should be seen increasing, however only to a degree. A somewhat inverse non-linear relationship is predicted, as cycling in either ends of extreme hot or cold weather is not something most people would do.

To determine how a correlation between two variables will be found, it must first be defined which correlation method is being used. For the current purpose of the investigation, the Pearson correlation coefficient was deemed suitable. The Pearson correlation focuses on the linear relationship and the measurements of that linear relationship's strength, between two variables. Hence, the strength of a correlation can only be deemed sufficient if the variables contain a linear relationship. [6] Furthermore, for a Pearson's correlation coefficient to be used, it is under assumption that the data must be normally distributed. The Pearson Correlation Coefficient follows the formula below:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2 n \sum y^2 - (\sum y)^2}}$$

With the r representing the Pearson's correlation coefficient, n being number of samples, x representing one sample, and y representing another sample. The range of the correlation coefficient covers the interval of all real numbers from -1 to 1. With -1 implying an inversely proportional relationship, 0 determining no linear relationship and 1 being a proportional relationship. To connect to the question, using the Pearson's correlation coefficient, it will be possible to measure the linear relationship between the various environmental factors and cyclist count. A dataset correlation matrix between the variables is presented below as a heat map.

There are certain ranges for the correlation coefficient, to summarise a poor correlation falls between (-)0.1 and (-)0.3, a moderate correlation between (-)0.3 and (-)0.5, and finally a strong correlation between (-)0.5 and (-)1.0. The strong correlation has a fairly large range, meaning that it is important to note that a correlation with 0.5 does not mean the same as a correlation of 0.8, for example. The same goes for other ranges as well. Therefore, it is important to interpret what the coefficients are and then draw conclusions.

In the case presented above, most coefficients seem to indicate a low negative linear relationship. To put into perspective the case mentioned in the above section, the correlation between windspeed and cyclist count is -0.23. This is not the same as the correlation between temperature and windspeed, which had -0.16. Although they are in the same "interval", these values represent different information. The -0.23 implies that windspeed and cyclist are more inversely proportional than windspeed and temperature. Further reasoning for such is provided in a section below.

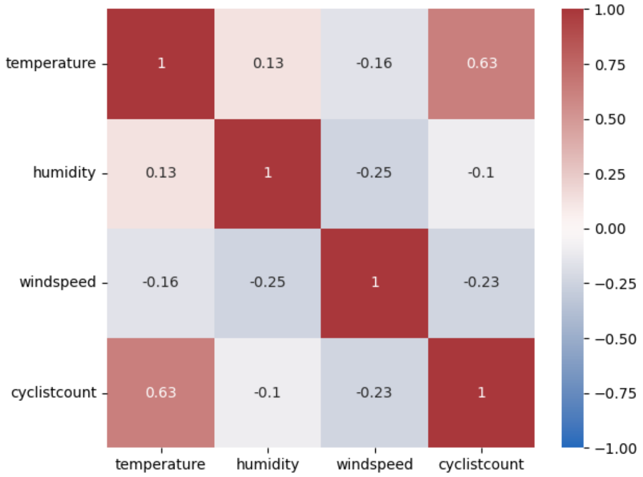

Figure 1: Heatmap of (Person's) correlation matrix consisting of environmental factors of the bike-sharing system dataset.

In regards to interpreting the results from the matrix, in general most variables do not contain a correlation above the low level. This is the case for environmental factors, with the highest being -0.25 between windspeed and humidity, however it cannot be concluded that windspeed and humidity are inversely proportional with these results. What it could imply, is that a further study should be conducted to look more into the relationship between windspeed and humidity. In terms of positive coefficients, there is one number that sticks out, and that is the correlation coefficient between temperature and cyclist count, standing at 0.63. Although it is on the lower end of the strong correlation, there still seems to exist some linear relationship, to an extent. Hence, it was decided to look specifically into the relationship between the two variables "cyclistcount" and "temperature".

A scatterplot was created, comparing the number of cyclists against the temperature to gain a better understanding of the two variables compared against each other. While, additionally checking to see if there are any outliers within the data itself which could be discarded for a more accurate investigation. The scatter plot is visible in the figure 2.

What was found was that in general, the data supports the premise that the number of cyclists increases at temperature also increases. With the all time high of cyclist count being at around the 25 degrees Celsius mark. Regarding outliers, there aren't any extreme ones, if there are any at all. Some points go slightly off at temperature around 25 and at the count of 2000 cyclists. Removing them however would not cause any great change as the size of the data is fairly large. In terms of patterns, the cyclist count does decrease at some point. After the temperature of 25 degrees, it is possible to see that it begins to fall back down. It means that cyclists prefer a temperature between 20 and 25 degrees, but tend to avoid more extreme temperatures because biking in much hotter or colder weather would be uncomfortable and more exhausting.

This data presented shows that there exists a proportional relationship between temperature and number of cyclists. The strength of the relationship is another factor. It must still be checked to what extent temperature can explain the number of cyclists, and this can

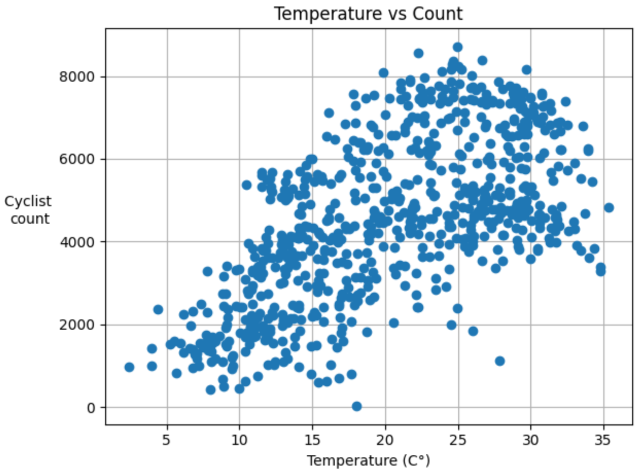

Figure 2: Scatterplot of all values comparing cyclist count and temperature in Celsius.

be done through the coefficient of determination. The coefficient of determination is the correlation coefficient to the power of 2. It is a measurement between the two variable's proportion of variance, in other words how one variable is capable of predicting the variance of the other. Ranging from 0% to 100%, the correlation coefficient between temperature and cyclist count was 0.63. Squaring this number results in roughly 0.4 or 40%. This in turn means that only 40% of the variation from temperature, is explainable in cyclist count. The majority of 60% is unexplainable. Hence, whether a concrete statement can be made about cyclist count increasing as temperature increases cannot be fully justified. Further investigations and data collection should be carried out in order to come to a more defined conclusion.

In terms of the hypothesis, and answering the research question provided, it was predicted mainly that between environmental factors themselves, there would be little to no visible correlation. The heatmap of the correlation matrix generally supports this part of the hypothesis, with windspeed specifically having a small negative correlation between it and cyclist count. For the temperature prediction, it seems that the investigation brought forth and extracted information that supported the idea of an inverse non-linear relationship being generated. The scatterplot showed that there was a substantial increase, and then a decrease. If more data were provided with more extreme temperatures, perhaps a more symmetric plot would be shown. To conclude and answer the question of "To what magnitude can correlations be seen existing between various environmental factors and cyclist count?"; it was demonstrated that the only environmental factor that had some form of strong correlation was temperature. Other factors were seen with some potential correlation, but had too little evidence from the data to have any sufficient conclusions. Hence, to only a low magnitude can it be said that some form of correlation exists between environmental factors and cyclist count, from the dataset explored.

3.2 Question 2

Hypothesis: Firstly, it must be established that seasons vary across the world, not only in the sense of inverted seasons (northern and southern hemisphere), but also that atmospheric conditions are dependent on the location of the recorded data. Hence, a generalisation cannot be done because of such variations, however given the fact that the recorded data came from the United States, it can be assumed that seasons and their corresponding atmospheric conditions are that of the northern hemisphere. The hypothesis came out at this point, is that seasons with comfier and days with nicer weather will see an increase in cyclist counts. It is predicted that summer will see larger increases, and decreases due to longer days, warmer temperatures and fairer weather. Then eventually these numbers will be seen decreasing, seeing the numbers rise and then fall because less people become available, and vacations come to an end. The general premise this discussion aims at, is that both season and weather type will have some form of impact on bike counts. To what degree this will occur, is what will be explored.

The first step taken towards answering the research question was to collect the relevant data, and organise it properly to prepare it to be used for graphical representation, so that it can be better interpreted. All fields except seasons, months, weather types and counts were removed, considering them irrelevant for the current section. It is important to see how the distribution of cyclists over the course of the year, which meant getting the sum of all cyclists within the months and plotting it on a line chart. Furthermore, considering that the recorded data spanned across 2 years, two plots were created for 2011 and 2012. This way, by having two lines, it is possible to see whether cyclist count can be seen increasing/decreasing during the same periods, which can then allow for confirmation of which seasons are most popular for cyclists.

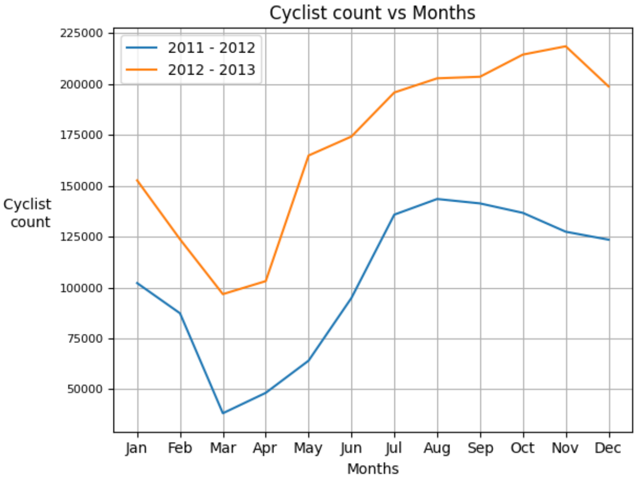

Figure 3: Line chart plotting distribution of cyclists against time (in months), showing similar patterns

It is important to note that the lines plotted, 2011 and 2012, should not be compared against each other. The question is not focused on the year, but instead how time in the year impacts cyclist numbers. Furthermore, seasons should not be tied to 3 months, as they are conventionally. For example, summer should not be read from 1st of June to 31st of August, but instead it spans from 21st of May to 20th of August. Taking these factors into account, it is possible to see that regardless of year, both lines see the largest rises from May to July, eventually falling off towards late autumn. The maximum's are found during August at 143512 (2011) and November at 218573 (2012). Interestingly, while the summer periods imply the largest gradients, the number continue to stay very high even after the season of summer comes to an end, implying that early summer sees some of the largest differences, therefore the cyclists will need to be tracked very carefully in order to meet demands. Whereas further down the line, towards autumn the numbers are rising more steadily and predicatively.

In 2011, the decrease from august (maximum) to December is only roughly 14% (down to 123511) and in 2012 the cyclist does not decrease during the summer period, and reaches its maximum in November, and only then is there a decrease of 9% (down to 198841). The original hypothesis thought that there would be increase for summer and decrease for autumn, however the data shown by the line chart indicates that while numbers rapidly grow as summer approaches, there is no sufficient evidence indicating any major decreases as was predicted. Taking the data collected and analysed, it can be stated that the findings support the premise that bike service demands are somewhat impacted by season, seeing numbers grow and fall throughout the year. An assumption to this result, was that the weather was often nicer during such time periods.

Correspondingly, it is only appropriate to follow up these findings by looking into how weather could have impacted the results. Seeing the results of the line charts, it would be meaningful to look into how much weather each season had experienced. By getting results as such it could be possible to interpret the weather, and see if there is some form of connection to the season with the highest number of cyclists. Therefore, to bring further statistical support to the claim to seasons, a multi-bar chart was created, which would also begin the investigation into weather impact on bike demands. The multi-bar chart can be seen as below in figure 4.

What can be seen from the chart, is that on most days, regard of season, it was clear or partly cloudy. There were no visible records of any days where it rained heavily or thunderstorms, and barely any days where there was light snow and rain. The data is fairly similar across all seasons, however autumn had the total highest number of recorded days where it was clear. A datapoint like this could be a possible explanation behind why there were high cyclist numbers during the autumn season. Since there are more days with more ideal weather, more people were active during that period. Regardless, even though autumn has the highest count, only the year 2012 displayed such results that were expected from weather. 2011 saw its highest peak during the summer season, although the numbers were still high, they were decreasing. An explanation for this could be that most of these clearer days were occurrent during the 2012 autumn, but there isn't enough sufficient evidence to support this claim. To further investigate how weather may have an effect on bike demand, box and whisker plots would be generated to see the distributions of the 3 occurrent weather types (Excluding heavy rain/thunderstorm).

To be noted from the chart, the red circle indicates outliers. The green diamond represents the mean. The orange line in the box is the median. With these pieces of information, it is possible to see skewness of each weather type, and data which are considered outliers. Some outliers existed in the "Clear-2012" category, implying that something must have caused a very small number of cyclists on those days, even though it was clear. It is definite that during rainy or light snowing days, the number of cyclists are usually lower (in that year), however there are days during mist/cloudy and clear days where the number of cyclists also dramatically decreases. Focusing on the boxes though, the clear weather category is always higher than the rest, meaning that third quartiles in those two categories are higher than 3rd quartiles in the other categories. It was however, interesting to see that the mist/cloudy categories maximum points were either similar or above the clear categories. They are less favourable conditions but people still chose to bike

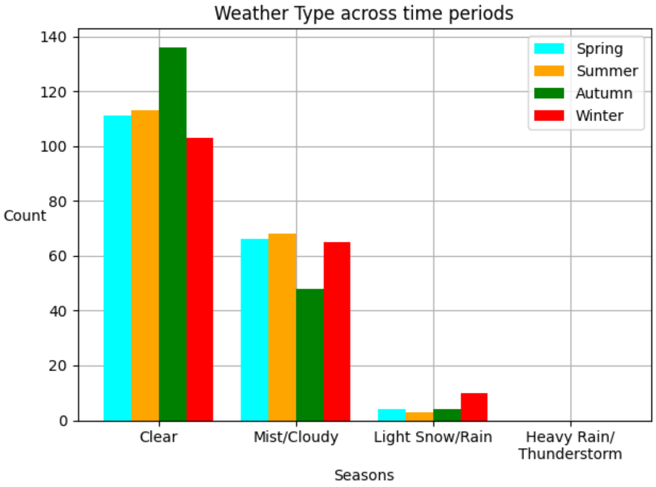

Figure 4: Multibar chart comparing time periods and recorded weather, showing what seasons had the most ideal weather and least ideal weather for cycling

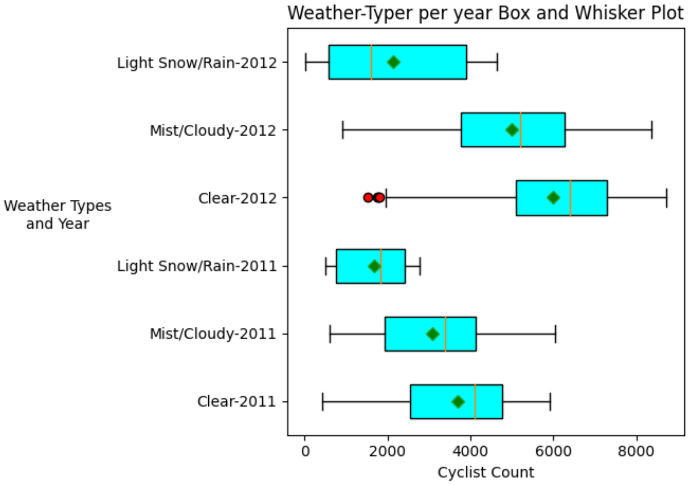

Figure 5: Distributions of weather-type vs cyclist count

on those days. Perhaps there was more of these records during holidays, or on weekends. Ultimately however, the medians and means were higher in the clear category implying that better conditions generally attract higher numbers of cyclists.

To summarise this section, the questions investigated what impact seasons and weather had on bike services demands. The theory's brought forth predicted that seasons with comfier atmospheric conditions, which included a higher number of nicer days, would be those with higher numbers. Hence, the season of summer was chosen as it had longer days, warmer temperatures and fairer weather. The prediction was to see an increase over summer then a decrease. The data collected and the line chart showed that although there was rapid growth at the start of summer, the numbers did not decrease as thought. Taking this into account, a multi-bar chart was implemented to find the season with the highest numbers of best days. The data revealed it to be autumn, which corresponds with the line chart as autumn saw very low decreases in 2011, but only increases during 2012. Weather typing was also investigated, which revealed that the number of cyclist are on average higher in the clear days section, with mist/cloudy being very close by, sometimes having even higher numbers. Taking all of this into account, and the data gathered, it can be stated that the season and weather display a visual effect on bike service demands to a large extent.

3.3 Question 3

Hypothesis: As was mentioned above, the prediction in regards to this question is then that the number of casual cyclists will be seen increasing, while the number of regular cyclists will decrease during business hours (especially in times such around 8-9am and 4-5pm when work/school starts and finishes). This will display some difference in bike usage throughout the early hours of the day. Furthermore, in the afternoon the number of casual cyclists should have some increase. With the reason being that regulars may be subscribed to that system because they need access to a bike in order to travel to/from, where a casual biker may get access to a bike for reasons of leisure, hence the predictions on the time frames.

To see whether a difference exists between casual and registered bike riders, it is important to get a visualisation of the data. To begin with, how do the two populations compare to each other, in terms of size. The reason for looking into this from the beginning is to make clear that any significant difference may not be a result of only the time, but also the actual number of bike riders. Therefore, the first part of this part to answering this question was to have a comparison of the sums of all hours of all days of casual riders and registered riders. The data had to be organised, and all numbers needed to be collected and separated into casual and registered. A pie chart was created to show the two populations, getting a clear image of the count of casual and registered cyclists.

As can be seen on the pie chart, overall, the number of recorded registered users, is vastly higher than the recorded number of casual cyclists. Therefore, this should be taken into account when discussing the results of the difference between casual and registered users in terms of hours. However, it should still be noted that this is not individual cyclists, but the number of recorded cyclists over the span of two years. It does not present information of how many registered users there are to casual cyclists, hence why the question is focused on the different hours throughout the day.

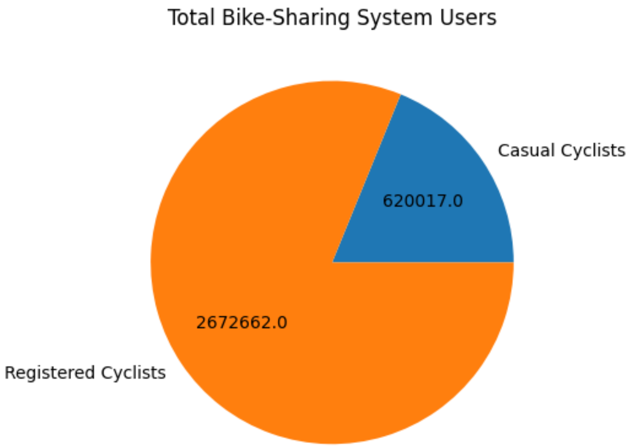

Figure 6: Pie chart showing recorded number of bike users

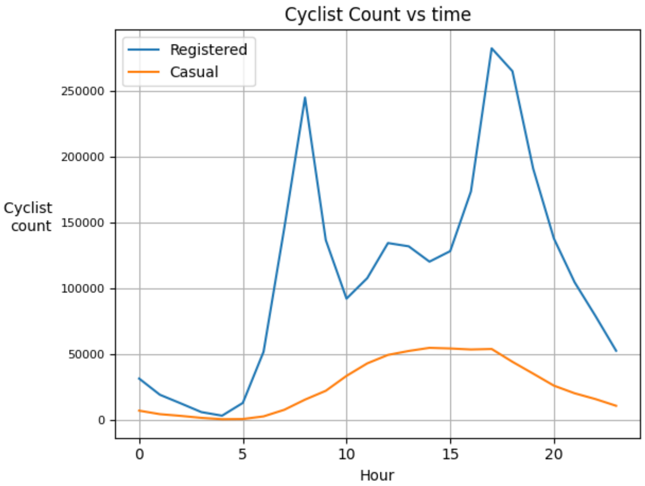

Figure 7: Line chart showing cyclist count over hours of the day

A line chart was chosen to visualise the number of cyclists over time through the day in figure 7. Two lines are plotted for casual and one for registered cyclists, in order to be able to better compare and visualise the difference between them. The numbers take into account both years of 2011 and 2012, and cyclist count is the sum for that hour across all records in the dataset.

From the graph it is possible to see that overall, the number of registered cyclists is always greater than the number of casual cyclists. This observation is supported by the pie chart, as the overall number is far greater, however there are some areas where the lines become very close, at the 4-hour mark. The difference in the early hours is most likely because most people would be asleep at this point, whereas business hours show a more prominent

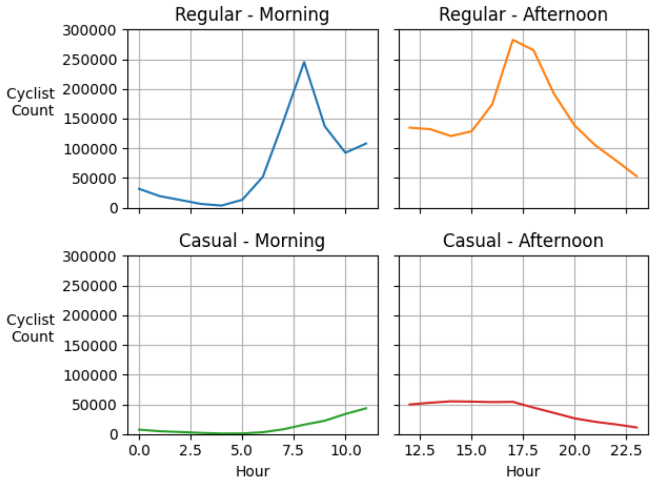

Figure 8: 4 subplots, splitting regular and casual cyclist count into two categories, morning (0-11) and afternoon (12-23)

difference. During early hours around 8-9am and 14-15 pm there are substantial increases in the registered cyclist count. These are the two peaks at which the number of cyclists is greatest, meaning that this distribution is a bimodal distribution. These are the local maximums in the data. These two peaks could even be separated and examined into two different groups, splitting the graph into early hours and late hours, and examine them separately in order to extract more specific information. As has been done in the figure in figure 8.

By splitting each group into two separate categories, the focus of comparing the two different groups may be easier. Additionally, the distribution of data may now be easier to compare. To have these distinctions allows to see exactly what activity points exist, which results in the skewness of the data. For regulars, in the morning the busiest hours are at 8, and in the afternoon its between 16 and 17. Whereas casuals do not have such high spikes, and show a steady incline and decline throughout the day. The casual-morning category, had a high positive skew at (1.223), similar to the regulars which had a high positive skew (1.037) as well. On the other spectrum, the casual-afternoon category had a moderate negative skew (-0.567), and the regular-afternoon group had almost opposite skewing (0.734). The information extracted from skewness visualises each group's major activity point, and indicating an interesting point that although the numbers between the group highly differ. The distributions for the first half of the day remain similar, while on the other end, the distributions remain opposite. Such observations build onto the premise that this section's research question explores.

What the line charts imply, is that most people who have a subscription with the bike-sharing system often use them to get to and from work or school. For casual however, the distribution is more evenly distributed, but it is still possible to see a negative skewing of the line. The casual number of cyclists is most prominent towards the middle/end of the day, where people often get off at work and can cycle for leisure. What can't be possible to determine from the graph alone, is whether this distribution is more impacted by weekdays, or by weekend days. What can be possible to see though, is that as the day progresses, and more people are free/awake, the number of casual bicycles can be seen increasing. Although the charts support the hypothesis of a difference between the two groups, further computations can be carried out in order to get more evidence to conclude on the difference between casual and regular cyclists.

To verify this difference mathematically, a two-sample test was done between the morning groups and the afternoon groups. Table 1 holds the values used for each test. The null hypothesis was setup to be that there is no difference between the means of regular cyclists and morning cyclists. The alternate hypothesis being that there will be a visible difference between the means. This test will be carried out with a significant level of 0.05/2 = 0.025. Since the sample size is sufficiently large, the z-score was used. The following formula was used to calculate the two z-scores:

Where t is the z score, the x represents the means; the s the standard deviation and the n the total number of samples.

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Table 1: Means (x), Standard deviations (x) and Sample (n) which will be used in the formula

Group	Mean	Standard Deviation	Sample
Casual - Morning	12135.25	14039.71	12
Regular - Morning	39532.83	16774.49	12
Casual - Afternoon	72350.42	75176.11	12
Regular - Afternoon	150371.4	68507.6	12

The data collected was then input into the formula. Two 2-Sample tests were carried out, one between casual and regular cyclists in the morning, and another between casual and regular cyclists in the afternoon. The critical range, which was based off of the significant level 0.05, was determined to be (+/-)1.960, converted from the z-score table. Anything below or above these values would result in the null hypothesis being rejected due to insufficient evidence. Another table is shown with the final outputs of the formula and the results underneath. C representing Causal, and R representing Regular. H_0 being the null hypothesis.

Table 2: Final results of the 2-sample statistical tests

Group	Z-Score	Critical Range	Result
C vs R Morning	4.34	(-)1.96	H_0 Rejected
C vs R Afternoon	2.66	(-)1.96	H_0 Rejected

Each 2-sample test revealed that the z-score was far higher than the critical range, meaning that the original premise of the means being the same is rejected. There is now mathematical evidence which supports a statistical significance on the effect, which is the difference between casual and regular cyclists resulting from the rejection of the null hypothesis.

To conclude on the data brought forth in this section, the research question focused on investigating whether a difference existed between casual and regular cyclists during the busiest hours of the day. It was predicted that through the earliest hours, there would be a difference between casual and regular users. Generally, the data supports the hypothesis presented. A pie chart first illustrated the recorded difference in the total number of observations between casual and registered users, implying some difference. Through the use of a line chart, the times of the day were included where vast differences were seen during the start and end of a work day, seeing some of the largest gaps and highest peaks in the regular cyclist counts. Splitting the graph further allowed for a more focused and separate investigation into the distributions and show how differences vary between morning and afternoon hours. To further confirm a difference mathematically, two 2-sample tests were carried out, with both resulting in a rejection of the null hypothesis. The means between regular and casual groups were different. Hence, taking all of the above mentioned into account, there is a substantial amount of data that supports the theory that a difference is implied and can be seen, visually and mathematically, between casual and registered cyclists during and outside of business/school hours. The difference investigated is statistically significant.

3.4 Limitations

This statistical report explored a dataset filled with information about a bike sharing service. Questions were presented, followed by data and analysis from which conclusions arose. This section takes into account possible limitations which may have had an impact on the results collected, and the conclusions drawn. Each question is presented with their respecitve limitation.

In question 1 the use of a heatmap and scatterplot allowed for there to be an exploration into the relationship between temperature and number of cyclists. Although it was found that some correlation existed between the two variables, there is some limitation to this piece of finding. To recall, the Pearson correlation was used, and it was mentioned that this was under the assumption of being used under normally distributed data. The dataset confirms that the environmental variables (temperature, humidity and windspeed) were normalized, hence this was fine for the Pearson correlation, however it was not confirmed for the cyclist count. Due to this uncertainty, it should have been more appropriate to use the Spearman rank correlation between temperature and cyclist count in order to draw upon a more accurate conclusion.

The conclusions that were drawn for the second question may also be subject to speculation. Something that is not taken into account is external events. Although there was strong data suggesting that weather and season do impact cyclist count, it cannot be ignored that the data could have been impacted by outside factors. In Washington D.C. 2012, there were several events ranging from sports events, food industry events to political/press events. More than 100 in 2012 [5]. With these being spread over the course of the year, and many taking several days the numbers should be subjected to investigation. Furthermore, the numbers in 2012 are vastly larger because of expansion to the capital bikeshare system, where 8 new systems were launched in Arlington. [1] These are some examples, however there could be others which have gone unnoticed by this study.

For question 3, the different types of biking groups were explored into more detail. The data looked into was split up into every hour of the day. The difference between the casual riders and registered riders was fairly large, and data showed how the activity levels increased and decreased through out the day, especially with the substantial increase during start and end of business hours for regulars. However, the data provided was not entirely accurate, and hence any conclusions drawn may be subject to some uncertainties. For example, hours 3 and 4 saw some of the lowest points on the graph, the observation of this error allows for this small number to now be explained. The reason for this being is that certain data was missing. In the dataset provided and explored in this report, there were rows missing and hence the number of hours were not the same (as they should have been). This left room for error with some numbers being lower than they should be. Table 3 presents how the numbers differ between hours.

Table 3: Different hours in the day and their number of rows

Hour	Count	Hour	Count
0	726	12	728
1	724	13	729
2	715	14	729
3	697	15	729
4	697	16	730
5	717	17	730
6	275	18	728
7	727	19	728
8	727	20	728
9	727	21	728
10	727	22	728
11	727	23	728

4 CONCLUSION

In essence, this report aimed at highlighting the growing importance of the bike-share service systems. The rapid increase in demands warranted an investigation into several factors, in order to draw insightful conclusions, which could be used to better understand, predict and apply newfound knowledge. The report used a dataset ranging from 2011 to 2012 located in Washington D.C, in the United States of America. To gain information from a range of areas, three questions were created which spanned across environmental factor impact, effect as a result of seasons-months and weather condition and finally difference between groups of cyclists in terms of time (hours) throughout the day. In the end it was found that a fair but not very strong correlation existed between temperature and numbers of cyclists, showing an environmental factor impact. Additionally, the seasons of summer and autumn were found to have the highest rises in cyclists, and number of cyclists, showing that seasons do fact make a difference. Finally, there was a significant difference between the casual and regular groups of cyclists, with regular groups often being more active during early and late hours when work/school begins and finishes. It was a clear observation that popularity of access to public bikes had growth, showing how crucial such analyses are required to paint a better image of the situation. However, considering possible limitations discussed there exists potential inaccuracies and uncertainties, which should not go unnoticed and therefore the conclusion is warranted to speculation. Regardless, bike demands will continue increasing, which paves ways for further analysis of the data in the future, presenting other findings and possible conclusions.

REFERENCES

[1] Capital Bikeshare. [n. d.]. About Capital Bikeshare. https://www.capitalbikeshare.com/about

[2] Christian Brand, Thomas Götschi, Evi Dons, Regine Gerike, Esther Anaya-Boig, Ione Avila-Palencia, Audrey de Nazelle, Mireia Gascon, Mailin Gaupp-Berghausen, Francesco Iacorossi, Sonja Kahlmeier, Luc Int Panis, Francesca Racioppi, David Rojas-Rueda, Arnout Standaert, Erik Stigell, Simona Sulikova, Sandra Wegener, and Mark J. Nieuwenhuijsen. 2021. The climate change mitigation impacts of active travel: Evidence from a longitudinal panel study in seven European cities. *Global Environmental Change* 67 (2021), 102224. https://doi.org/10.1016/j.gloenvcha.2021.102224

[3] Matthew Norris Edward McMahon and Rachel MacCleery. 2016. Active-Transportation-and-Real-Estate-The-Next-Frontier.

[4] Hadi Fanaee-T and Joao Gama. 2013. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence* (2013), 1–15. https://doi.org/10.1007/s13748-013-0040-3

[5] Beth Kormanik. 2012. Washington's Top 100 Events 2012. https://www.bizbash.com/bizbash-lists/top-100-events/top-list/13230517/washingtons-top-100-events-2012

[6] David M. Lane, Mikki Hebl David Scott, Dan Osherson Rudy Guerra, and Heidi Zimmer. 2003. *Introduction To Statistics*. David M. Lane.

[7] P Oja, Sylvia Titze, Adrian Bauman, Bas De Geus, Patricia Krenn, Bill Reger-Nash, and T Kohlberger. 2011. Health benefits of cycling: A systematic review. *Scandinavian journal of medicine I& science in sports* 21 (04 2011), 496–509. https://doi.org/10.1111/j.1600-0838.2011.01299.x