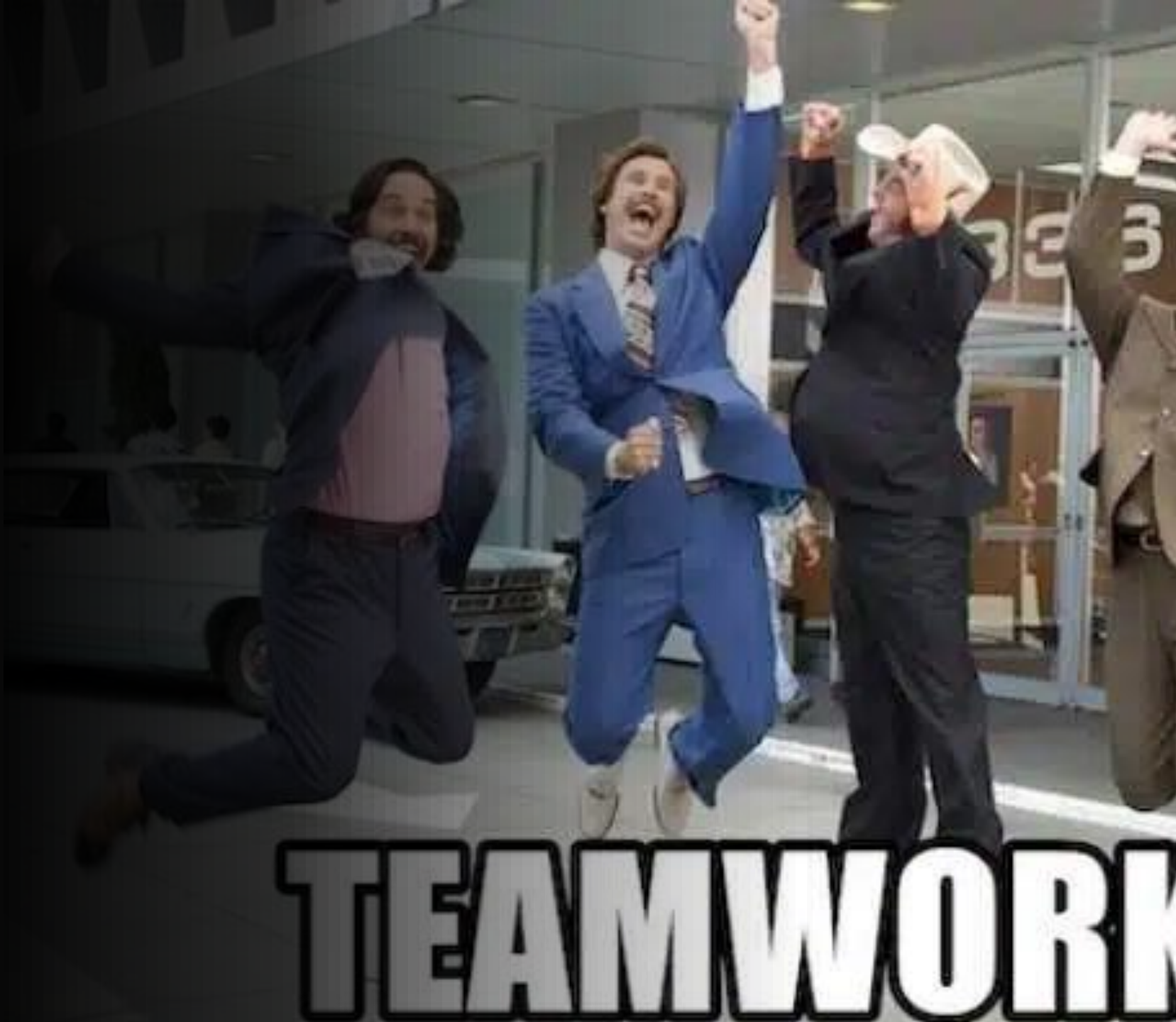




Roadmap for The Analytics Alchemists

- Coming together is a beginning, staying together is progress, and working together is success.

– Henry Ford





Paris Agreement

Article 2


1. This Agreement, in enhancing the implementation of the Convention, including its objective, aims to strengthen the global response to the threat of climate change, in the context of sustainable development and efforts to eradicate poverty, including by:

(a) Holding the increase in the global average temperature to well below 2°C above pre-industrial levels and pursuing efforts to limit the temperature increase to 1.5°C above pre-industrial levels, recognizing that this would significantly reduce the risks and impacts of climate change;

(b) Increasing the ability to adapt to the adverse impacts of climate change and foster climate resilience and low greenhouse gas emissions development, in a manner that does not threaten food production; and

(c) Making finance flows consistent with a pathway towards low greenhouse gas emissions and climate-resilient development.

2. This Agreement will be implemented to reflect equity and the principle of common but differentiated responsibilities and respective capabilities, in the light of different national circumstances.



Paris Agreement

- What were mean temperature before the industrial period?
- How is mean temperature evolved?
- Where are temperature heading?
- How is temperature changing in different continents / parts of the world?

Article 2


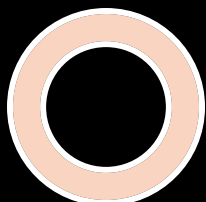

1. This Agreement, in enhancing the implementation of the Convention, including its objective, aims to strengthen the global response to the threat of climate change, in the context of sustainable development and efforts to eradicate poverty, including by:

(a) Holding the increase in the global average temperature to well below 2°C above pre-industrial levels and pursuing efforts to limit the temperature increase to 1.5°C above pre-industrial levels, recognizing that this would significantly reduce the risks and impacts of climate change;


(b) Increasing the ability to adapt to the adverse impacts of climate change and foster climate resilience and low greenhouse gas emissions development, in a manner that does not threaten food production; and

(c) Making finance flows consistent with a pathway towards low greenhouse gas emissions and climate-resilient development.

2. This Agreement will be implemented to reflect equity and the principle of common but differentiated responsibilities and respective capabilities, in the light of different national circumstances.



Analysis of greenhouse gas emission trends and national contributions

- A data scientist can collect and analyze data on greenhouse gas emissions from different countries, based on Nationally Determined Contributions (NDCs) reported in accordance with Article 4 of the Paris Agreement. This may involve:
 - Collecting data from public databases such as the UNFCCC (United Nations Framework Convention on Climate Change) data repository, including national inventories of anthropogenic emissions (Article 13, paragraph 7).
 - Using statistical models and machine learning to identify trends in emissions of CO₂, methane (CH₄), nitrous oxide (N₂O) and other greenhouse gases over time.
 - Comparing emission trends with the Paris Agreement's goal of limiting global warming to well below 2°C, and pursuing efforts to 1.5°C (Article 2, paragraph 1).
 - Evaluating whether countries' contributions represent a progression over time, as required by Article 3, and identifying gaps or deficiencies in the level of ambition.
 - This work can provide insight into which countries need additional support (e.g. developing countries, Article 9) and how global emissions can reach balance between emissions and removals of greenhouse gases this century (Article 4, paragraph 1).
- 

Article 4

1. In order to achieve the long-term temperature goal set out in Article 2, Parties aim to reach global peaking of greenhouse gas emissions as soon as possible, recognizing that peaking will take longer for developing country Parties, and to undertake rapid reductions thereafter in accordance with best available science, so as to achieve a balance between anthropogenic emissions by sources and removals by sinks of greenhouse gases in the second half of this century, on the basis of equity, and in the context of sustainable development and efforts to eradicate poverty.

Greenhouse gases

- What is the current level of anthropogenic greenhouse gas emissions?
- What is the correlation between greenhouse gases and surface temperature?
- How are greenhouse gases affecting forestation/deforestation (carbon flux)?
- Progress of developed and low-developed countries.

Reduce greenhouse gases

- Can we measure sinks in greenhouse gases?
- Developed and low-developed countries.
- Carbon flux (life cycle).

Article 5

1. Parties should take action to conserve and enhance, as appropriate, sinks and reservoirs of greenhouse gases as referred to in Article 4, paragraph 1 (d), of the Convention, including forests.

2. Parties are encouraged to take action to implement and support, including through results-based payments, the existing framework as set out in related guidance and decisions already agreed under the Convention for: policy approaches and positive incentives for activities relating to reducing emissions from deforestation and forest degradation, and the role of conservation, sustainable management of forests and enhancement of forest carbon stocks in developing countries; and alternative policy approaches, such as joint mitigation and adaptation approaches for the integral and sustainable management of forests, while reaffirming the importance of incentivizing, as appropriate, non-carbon benefits associated with such approaches.

Article 9

1. Developed country Parties shall provide financial resources to assist developing country Parties with respect to both mitigation and adaptation in continuation of their existing obligations under the Convention.

2. Other Parties are encouraged to provide or continue to provide such support voluntarily.

Financial investments

- Overview of investments.
- Efficiency of investments?

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 CO₂ emissions. Saving 1 trip a day, for at least 200 days a year reduces half a tonne of CO₂ 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:

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

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

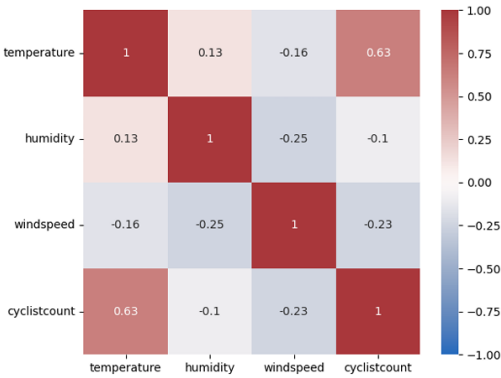


Figure 1: Heatmap of (Person's) correlation matrix consisting of environmental factors of the bike-sharing system dataset.

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 as 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,

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

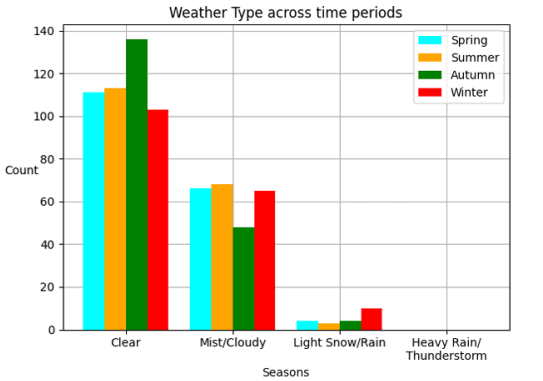


Figure 4: Multibar chart comparing time periods and recorded weather, showing what seasons had the most ideal weather and least ideal weather for cycling

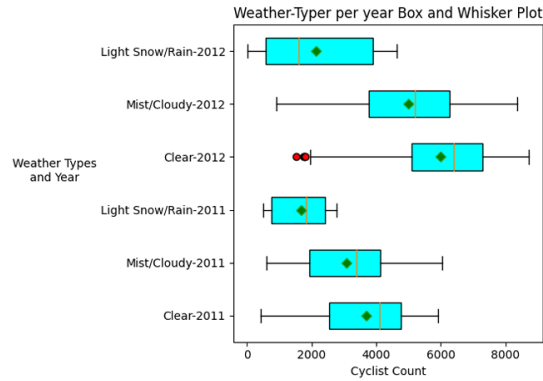


Figure 5: Distributions of weather-type vs cyclist count

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

Bike Sharing Services: A statistical analysis report

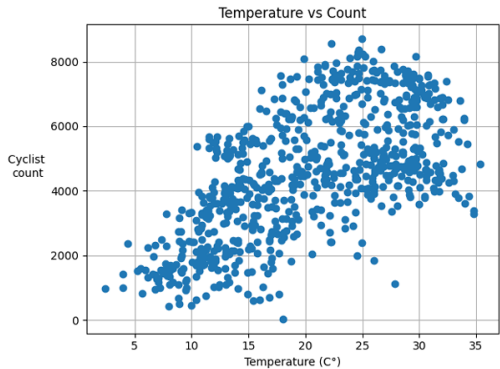


Figure 2: Scatterplot of all values comparing cyclist count and temperature in Celsius.

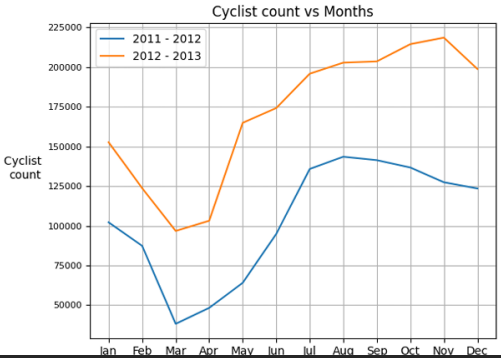
be done through the coefficient of determination. The coefficient of determination is the correlation coefficient to the power of 2. It is a measurement between the two variable's proportion of variance, in other words how one variable is capable of predicting the variance of the other. Ranging from 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

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.





YOU DID A GOOD JOB.

What needs to
be done

**FALSE. YOU DID AN AWESOME
JOB**



We need to:



To analyze datasets !!!



Write our findings !!!



Handle sources and Overleaf



Prepare presentation (45% of accumulated score)



How will you
benefit the
team?

