2022

Maria Bocharova

CSE 696

7/30/2022

Bike-sharing demand vs CO2 emission reduction; can bike sharing bring a cleaner future ?



**Table of Contents**

* **Introduction**
* Background of the study
* Statement of the problem
* Objective/ aim of the project
* Significance of the project
* **System Requirements**
* System specification
* Hardware/software overview
* **Method**
* Detailed implementation
* Data collection/calculations
* Procedures
* Hardware/software implementation
* **Analysis**
* Evaluation and analysis of the results
* **Discussion**
* Results
* Contribution of the project to the world
* Next steps
* Improvements
* Flaws
* **List of Figures**
* **References**

**INTRODUCTION**

With the rapid growth of major cities and an increase in gas prices bicycles are becoming one of the primary sources of transportation for many citizens. Bike sharing is becoming more widely accessible, and preferable to users over personal bikes. A bike-sharing system is a public transport service that makes bicycles available for sharing to individuals on a short-term basis for a small fee or sometimes free of charge. Most bike-sharing systems allow customers to get and return bikes at the docs or other designated locations. According to world bike sharing statistics revenue in the bike-sharing segment is projected to reach $7.63bn in 2022 from $3.43bn in 2019.

Bike sharing is not only a convenient way of transportation for many users; bicycles also use minimal fossil fuel and are generally a pollution-free mode of transport. They are also requiring a minimal number of materials used to produce and thus can help reduce the need to build, service, and dispose of cars or other heavy modes of transportation. It is also a potential way to drastically reduce the release of CO2 in the air - the main gas produced by cars. Automobiles also produce methane (CH4) and nitrous oxide (N2O); all those substances are known as greenhouse gases. They are trapping heat and make the planet generally warmer. According to EPA (the United States Environmental Protection Agency) transportation contributes 27% of total greenhouse gas emissions in the world.

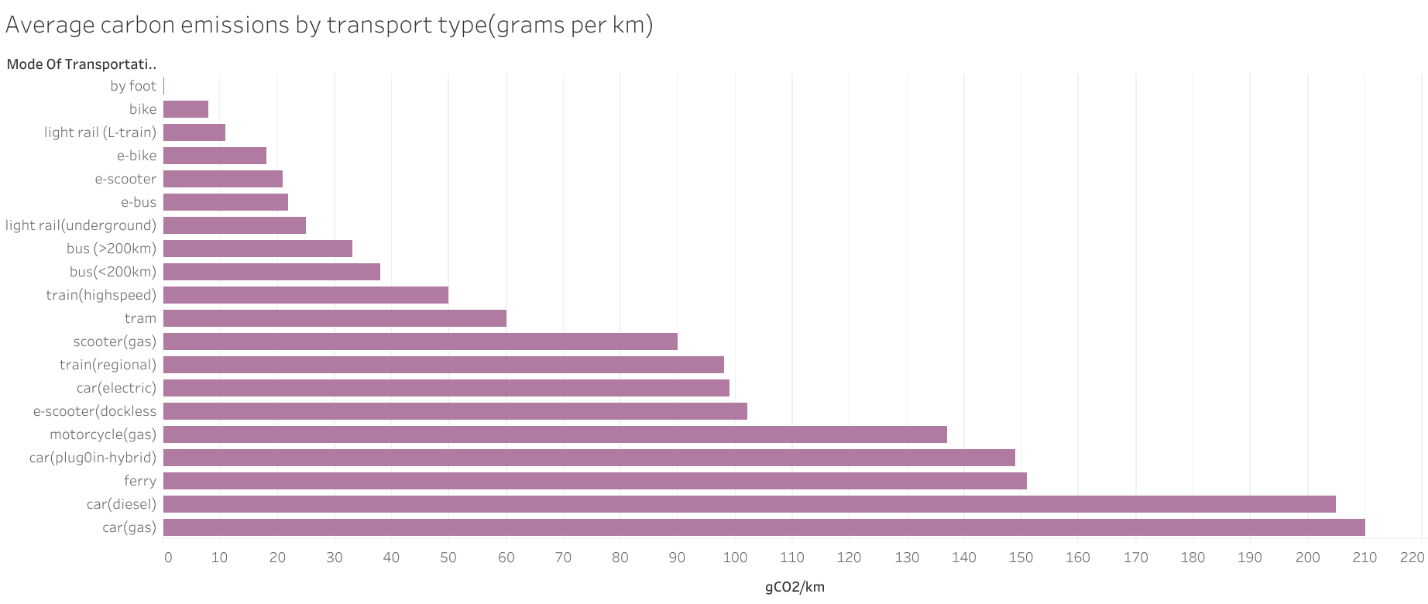


Figure 1 Average Carbon Emissions by Transportation Type (grams per kg)

On the graph (Figure 1) above average modes of transportation in the big cities can be observed; starting from the most eco-friendly one (by foot) to the worst (car on gasoline). However, what is an eco-friendly mode of transportation? Not only the amount of carbon emitted during the use of transportation should be considered but also, how much effort, materials, and time it is taking to produce, transport, and store that means of transportation. For instance, for an electric car, around 80% of its emissions come from manufacturing and disposal, while for a regular car only around 25% comes from manufacturing and disposal and the rest from direct and indirect operations. For the bicycle on the other hand most of its emissions are the consequences of manufacturing. And thus bicycles become main focus of this research.

The rapid increase of bike sharing is a signal for a potential effect on greenhouse gas emissions in major cities around the globe. The goal of this project is to analyze whether or not using bike share in major US cities has an effect on CO2 reductio (or increase). During the research additional factors that contributed to the overall emission were discovered and had to be added to proceed with the final outcomes. Particularly those factors are variations in diets among cyclists. It is an important variable as some bicycle users, who, for example, have a meat-heavy diet could contribute more to pollution than not.

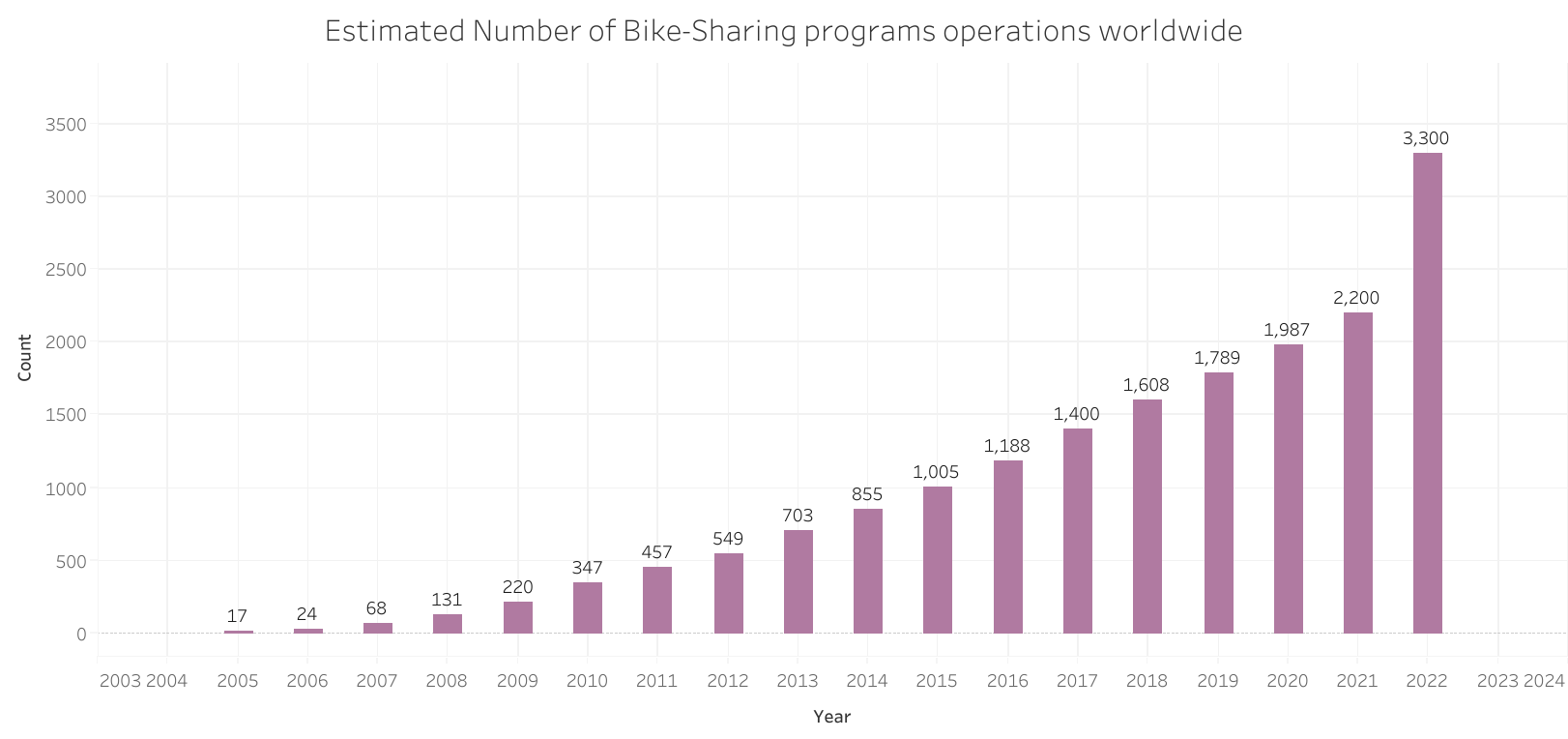


Figure 2 Estimated Number of Bike Sharing Programs Around the World (c.2022)

Several USA cities were chosen as a target for identifying the significance of bike sharing systems in CO2 reduction: Boston, Austin, New York, Los Angeles, and Chicago. All the cities were chosen based on their territorial location and access to public bike-sharing programs.

**SYSTEM REQUIREMENTS**

This analysis was done using a minimum requirement of Python 3.10. It was done on Windows 10 machine using Jupyter Notebooks, Tableau, and RStudio.

Required packages for loading the data are:

* Pandas’ version 1.3.4
* Zip file version 3.2

Require packages for data preprocessing:

* Pandas’ version 1.3.4
* Seaborn version 0.11.2
* Matplotlib version 3.4.3

Required packages for data analysis/forecasting:

* Panda’s version 1.3.4
* Matplotlib version 3.4.3
* Seaborn version 0.11.2
* Sklearn version 1.0.2
* Numpy version 1.20.3
* Math
* Statsmodels version 0.12.2
* Plotly version 5.9.0

All code was developed with and compiled in a jupyter notebook. The code is built to process simple CSV files as input. Original data was scraped from the internet into several tables for different cities, months, and years. Data could vary from city to city, however, some of the columns had to be universal, such as city, date, distance, and user type, therefore, data was collected accordingly. Optional columns included gender, starting, and ending locations names were present but not used for preprocessing and further analysis. Additional columns were calculated and added to the database later; those columns defined estimates on how much emission is produced by cyclists following various food diets, alongside emissions produced by cars during the same amount of mileage traveled.

**METHOD**

Data is a core entity of this project; therefore, data collection was a thorough process. The main objective was to gather information-rich and reliable data. At first, it was planned to obtain an existing dataset, however, after some research, none of the available sets satisfied the required criteria.

Only the CVS file formatted accordingly can be used for this program. As it is more of an exploratory and predictive analysis, the file must have identical formatting as the one used for the original research. The file should be single-headed with at least three columns: user type, time/date, and duration. However, as data was collected from multiple sources, it would be hard to define more particular requirements for the file, and the user should adjust based on their needs.

The data for the original file was collected from multiple data sources\*. The goal was to collect bike-sharing information from five major US cities. All these sources are government websites that obtain transportation data about their citizen from third-party groups. At the beginning of data collection, it was determined what variables to look for: the main variables were dates, city, duration, and type of user. All five websites that were used had this information available in various formats and forms. All the data was continuous, therefore, timeframe constraints had to be determined. It was decided to use a time period from the beginning (January) of 2019 to the end(December) of 2021. This period was optimal as it included time before, during, and after the global pandemic (“CDC Museum COVID-19 Timeline”).

Once the data was collected, the findings were organized. Data obtained from two cities (NYC, and Chicago) was dropped due to the limited amount of computer memory available – the data was too heavy to process.

The first step of data preprocessing was to scan and conclude what data will benefit the project and should be kept and what data will be dropped. Original data was obtained in zip files, after using the zip file library to unzip the package into one dataset, large numbers of extra columns were created which increased the complexity of the dataset. All the columns were of various data types and formats; they were normalized and scaled using NumPy and pandas.

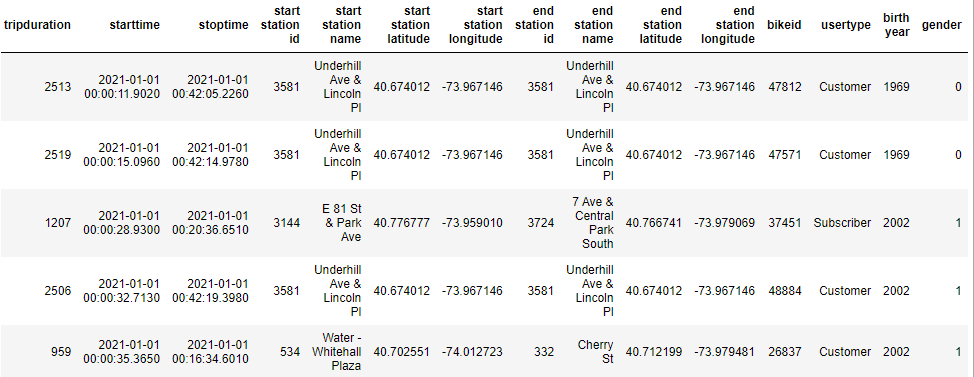


Figure 3 Example of original data for single month entry and single city (New York)

Obtained data were merged into three different datasets based on the city. Each dataset contained the same main columns (‘date’, ‘duration’, ‘user type’), in addition to various other columns such as ‘gender’ in one dataset and starting/ending ride station names in another. As data were collected from unrelated sources, the terminology was different for all three tables and was changed as well. The user type, for instance, varied from table to table from military/student to single day pass, to up to 30 different user types for one city. They were split into two categories: Customer and Subscriber and normalized according to that. The sets were data heavy, therefore, the random selection had to be done. Originally each dataset contained more than a few million entries and was reduced to around ~360K rows each, which totals slightly over ten thousand entries for each month. After that, the data was merged into one dataset.

Original data contained starting and ending dates and times; this column was split into two for time only and date only. The time column was used to calculate how long each trip was. After that trips that a less than 6 min were dropped alongside trips that were over 8 hours or 499 min. The next was to calculate the estimated distance it takes to travel in those intervals of time. According to the Urban Cycling Organization, the average speed for a regular city commuter is around 12 mph. Therefore, the speed of 12mph was estimated for every entry in the table. Then as average speed and time were known, distance traveled in miles was also calculated:

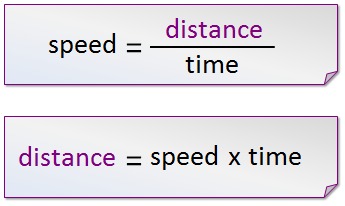


Figure 4 Distance Formula

As it was mentioned in the introduction, biking has a possible greater impact on the climate than it was thought of but depending on the diet you eat. For instance, a person consuming more meat products may burn more calories by biking but this lead to similar emissions as a fuel-efficient car (land use, production, and all other factors considered).

It was first estimated how much energy is needed to travel by car and by bicycle. Preliminary data showed that the average USA car fuel efficiency is 24.2 mpg (“Alternative Fuels Data Center: Maps and Data - Average Fuel Economy by Major Vehicle Category”). 24.2 miles per gallon is equivalent to 9.719 liters per 100 kilometers (62.1371 miles). There is 25MJ in 1L, therefore, it is equivalent to (9.719\*25) MJ / 100 km. It totaled around 2.43 MJ/km.

In comparison, the research from Geus found that cycling requires around 25 kcal/km (or around 40 kcal/mile) above basal metabolism (number of calories burnt when the body performs basic life-sustaining activities). It is equivalent to .1046 MJ/ km (or 0.16736 MJ/mile), which is 23 times less energy than required by a car. However, this is to be expected as cars are heavier by weight than bicycles: 4,156 lbs. vs 18 lbs.

Global energy consumption continues to grow, and not all sources have a similar impact on climate. For this particular research, the main focus is on Carbon Dioxide or CO2 gas as it is one of the main byproducts of energy production – it is known as a greenhouse gas. There are many more gases (Figure 5) that are in the same group, however, they also are usually mentioned as CO2 equivalents. Their individual fractions are drastically smaller; therefore, it is hard and irrelevant to this research to determine their exact values.

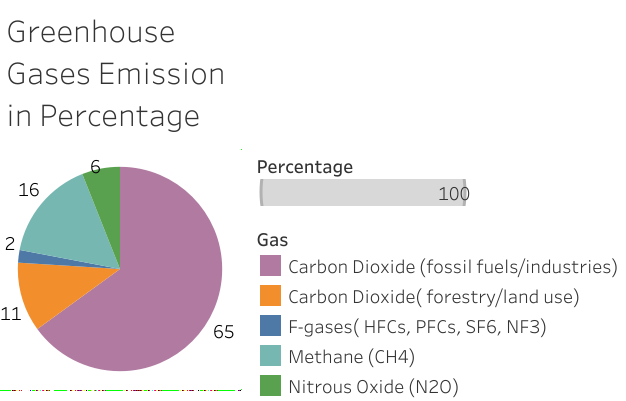
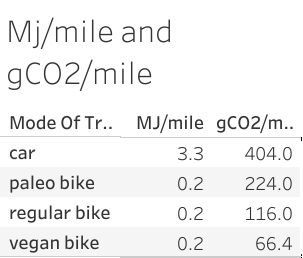


Figure 5 World Greenhouse Gasses Emissions in percentage

The next step was to calculate how much CO2 per mile is emitted by cars and bicycles. According to the United States Environmental Protection Agency, a typical car has around 404 grams of CO2 per mile driven. Similar calculations can be done for cyclists with various diets. The average man consumes between 2,400 and 3,000 calories a day while woman between 1,800 and 2,400 per day. It is on average 2700 kcal a day. According to the research average person produces around 2.9 g CO2/kcal following a regular diet and 1.66 g CO2/kcal following a vegan (no animal products consumed). (Scarborough). Meat-based diet was estimated using statistics from ‘Our World in Data’ by adjusting the kcal intake only from animal products, which was equal to around 5.6 g CO2/kcal for a rate of 2700 kcal a day.

The average bicyclist produces 0.16736 MJ/mile which is equivalent to 40 gCO2/mile - it is already ten times less than produced by the car. However, various diets should be taken into the consideration. As mentioned above regular diet is equivalent to 2.9 gCO2/kcal or 116 gCO2/mile, a vegan diet is 1.66gCO2/kcal or 66.4 gCO2/mile, and 5.6 gCO2/kcal for a meat diet which is equivalent to 224 gCO2/mile.



The calculations were applied to each entry in the table based on the miles traveled to get the total CO2 emission per trip. This data was later used to analyze the correlation between diet and CO2 emissions. Diet, however, is not a controllable variable and the main analysis was focused on the number of users per each time period. The number of users is a defining factor as it directly correlates to a reduction of automobile use for daily commutes and leisure. To determine the growth or decline of cycling commuters Autoregressive integrated moving average was used. It is a time series analysis of an autoregressive moving average model, meaning it measures the time series of a constantly evolving dataset. This model was chosen as it is proven to have accurate predictions for future outcomes using past values.

As the first step for ARIMA (Autoregressive integrated moving average) dates were changed to a python DateTime format. After those dates were reformatted to ‘%Y-%m-%d’ form and set as indexes for a dataset – indexes are needed to represent relation to time. All the time series have three main components: trend (upward/downward movement), seasonality (seasonal variance), and noise (spikes at random intervals). They all are useful for predicting forecasting, however, all three of them may not be present for every dataset.

The next step was to make sure that data is in stationary form (Figure 6) – meaning the mean stays the same throughout time, as usually for time series data is continuous as the mean fluctuates.

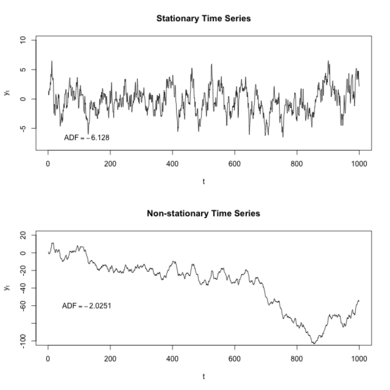


Figure 6 Stationary and Non-Stationary Time Series Example Graphs

The next step was to analyze variance(Figure 7) – it refers to how spread the dataset is and, also, should not be a function of time and stay the same throughout the dataset.

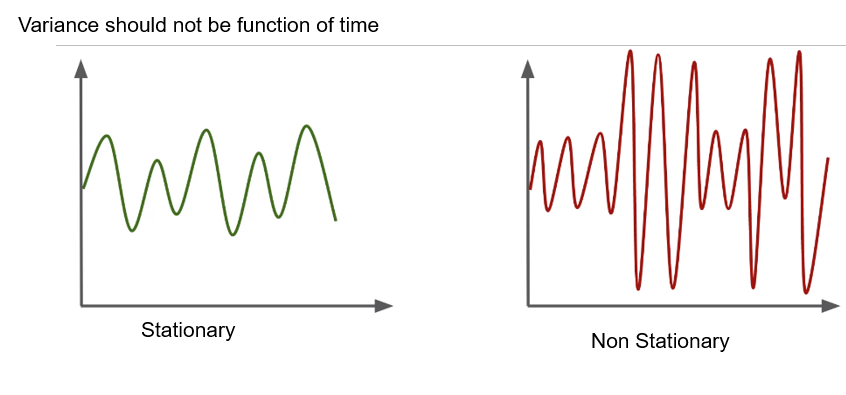


Figure 7 Stationary and Non-Stationary Variance example graphs

And the final step was to make sure covariance (a measure of joint variability of two random variables) is also not a function of time. It should remain stationary as well. Data being stationary is very important as it allows us to predict future behaviors based on past behaviors.

Next, all the important libraries were imported such as NumPy, pandas, matplotlib, and statsmodels which provide a wide variety of tools that help working with time series. The main data was then preprocessed again to fit the needs of the program. The date was limited to month and year only, and the total user count was calculated for each month accordingly. Start\_date is used as an index.

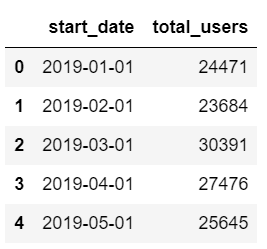


Figure 8 Reduced Dataset Head Example

As can be observed in the graph below (Figure 9) there is a trend component in the dataset but not very consistent.

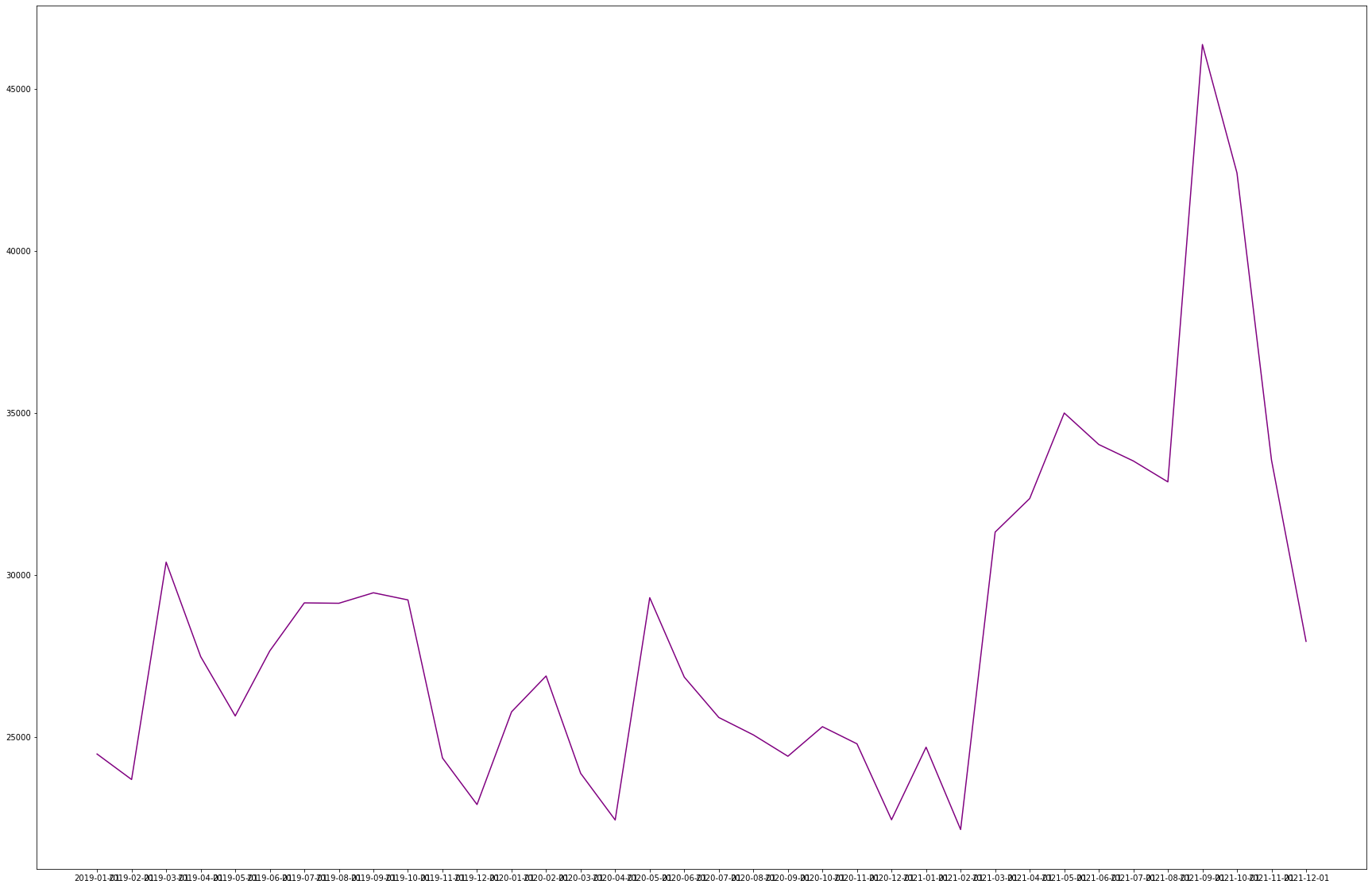


Figure 9 Total User Count Stationary Data with a Trend Component. The Trend Component is Not Distributed Evenly

The next step is to make sure that the data is stationary. There are several methods, but two primary methods were used in this project: rolling statistics and the augmented Dickey-Fuller test. For rolling statistics mean and standard deviation must be plotted, and they must remain constant to prove that data is stationary. The second method is using a p-value (a measure used to validate a hypothesis against the observed data) as ADF is a statistical significance test. The ADF is a ‘Unit Root Test’ meaning a unit root is existing in a time series when the value of alpha is equal to 1 in the:

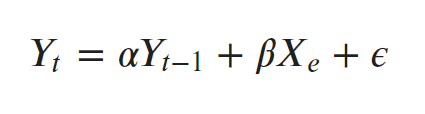


Figure 10 Unit Root Test Formula

Where Yt is the time series value at time t and Xe (an exogenous variable which is a separate variable in a time series). In simple words, if there is a unit root, the time series is not stationary. Dickey-Filler test tests the null hypothesis that alpha equals 1, and if the hypothesis is not rejected the series are considered nonstationary. The Augmented Dickey-Fuller test involves high order of the regressive process in the model:

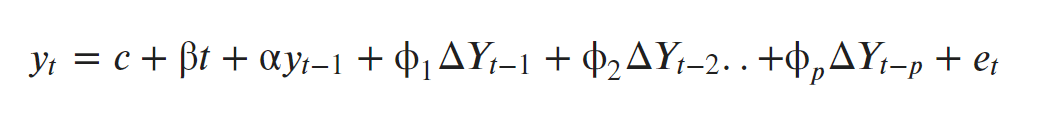


Figure 11 Augmented Dickey-Fuller Test Formula

This equation is partially similar to the unit root test but with an addition of extra variables. Hence, as the null hypothesis assumes the presence of the unit root, the p-value plays a significant role and must be below 0.05 to consider it valuable, and reject the hypothesis, proving that data is stationary. Also, if critical values were to be closer to 1%,5%, and 10% that means that confidence intervals are critically close to the ADF results, and data is again stationary.

ARIMA consists of two parts AR or Auto Regressive model and MA or Moving Average model. AR is based on the functionality that past values have an effect on the current values, and as long as an assumption is valid, a linear regression model can be built to predict dependable values given a linear combination of previous timestamps. MA on the other hand assumes that the current value is dependable on the error terms of the previous timestamps. The error term is a difference between the expected value at a particular time and the value that was factually observed. ARMA model then combines them both.

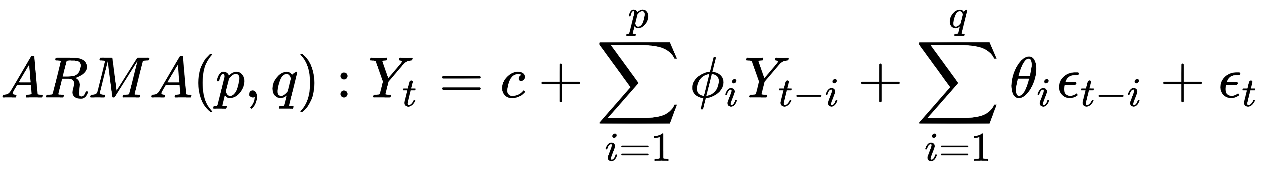


Figure 12 ARMA Model Formula

In the formula above the first part is AR and the second MA or error terms formula – the ARMA is simply an addition of the two of them.

ARIMA is the updated version of the ARMA – it has the same components with an addition of the differencing. The differencing is a transformation of the time series to a new time series where the values are differences between the consecutive values.

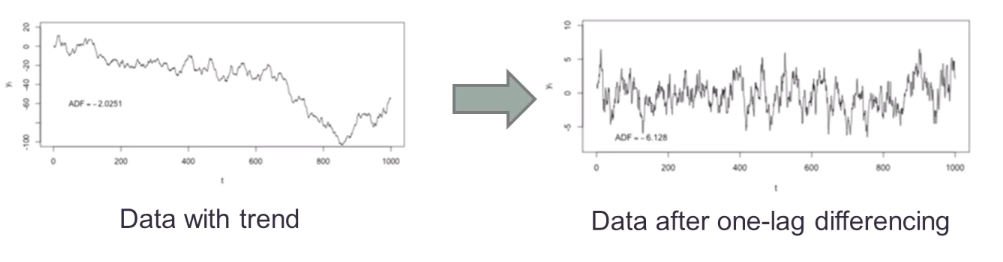


Figure 13 Example of Data with Trend Component (left) and Differencing Component (right)

It works by subtracting the value of the preceding day from the current value. As an example, first-order differencing is using linear terms and have a form of Xi = Yi-Yi-1, while second-order differencing is quadratic and has a form of Xi =(Yi-Yi-1) -(Yi-1-Yi-2) – preceding day is subtracted from a current day, and the preceding day is subtracted by the day before that, and so on. On the graph above (Figure 13) it can be observed why it is important to use differencing as it removed the trend and creates more evenly distributed data.

ARIMA model is run with three main parameters - p, d, and q. P is for the number of autoregressive terms (order or how many timestamps are considered). D is for the number of nonseasonal differences. Q is for the number of moving average terms (is also a number of terms to include in the equation). The auto Correlation function (ACF) is used to identify the number of occurring MA terms and the Partial Auto Correlation Function (PACF) is used to identify a number of occurring terms in the AR model.

Both PACF and ACF are theoretical statistical models such as variance or expected value that are used for random variables, PACF and ACF are used in the same manner but for time series. PACF for AR eliminates past partial autocorrelations and makes them equal to 0, and a number of non-zero partial autocorrelation gives an order of the AR model. Order is a lag–fixed amount of passing time that is considered as a term period for the model.

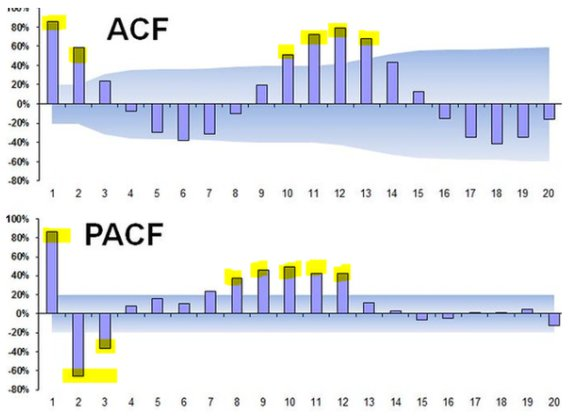


Figure 14 ACF (top) and PACF (bottom) graphs examples

The horizontal marked area on the Fugure 14 corresponds to the threshold of significant values, and all the spikes that are above it are considered to be valuable terms. For the example above: ACF has 6 terms total and PACF has 9 terms total.

The next step is to train the model and predict the fitted values, alongside the forecasting prediction for the future. ARMA model was also used for comparison of the results.

For ARMA the dataset was split into test and train sets using the python SARIMAX package. It is trained on quantitating stationary variables and in the case of this dataset, it is the total user count per month. The same term order was used – 2,1,2. The model then was fit, and a forecast was created as well.

Chart, line chart

Description automatically generatedChart

Description automatically generated

Figure 15 ARMA (left) and ARIMA (right) Graphs for Comparison

ARMA model explains the relationships of time series with both moving average (random noise) and autorepression (values at previous timestamps). The presence of a linear trend in prediction and a smaller value of RMSE (Root Mean Squared Error – the square root of Mean Squared Error (MSE)) will signify a more accurate model.

**Hardware Implementation**

There are no requirements for hardware for this project. For instance, it can be done on any machine with 2 – 4GB ram and ability to download the required software packages. The data was, however, collected using hardware installed on the operating bicycles. Those bikes are using embedded bike tracking tech such as GPS microchips to record the location, distance, mileage, and time, and send it to the user’s phone, alongside an online database.

**Software Implementation**

Structure

* chicago\_data.ipynb (identical concept and procedures for other cities' data):

contains necessary lines of codes to load, and normalize data for individual cities.

* load\_preprocess\_final.ipynb:

contains functions that preprocess final combined data

* seaborn\_graph.ipynb:

contains functions that help reduce and visualize the original dataset

* prediction\_graphs.ipynb:

contains functions that help create exploratory analysis and predictions and visualize them

* prediction\_graphs\_2.ipynb:

contains functions that help create exploratory analysis and predictions and visualize them

* graphs.ipynb :

contains functions for various diets comparative analysis

* arima\_arma\_models.ipynb:

contains function for forecasting ARIMA and ARMA models

**ANALYSIS**

The collected data was analyzed using Python via Jupyter Notebook and Tableau to determine the average length of the trip, amounts of CO2 released based on various diets and trip length, and to construct predictive analysis for future city bicycle usage - whether or not bike sharing can have an effect on a constantly changing climate.

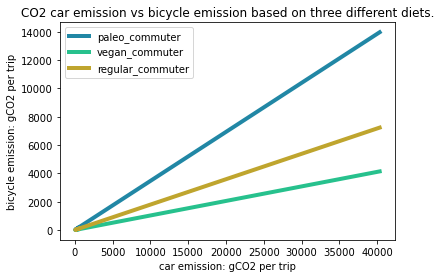


Figure 16 Three different diets (regular, plant-based (vegan), meat-based (paleo)) and their relative CO2 emissions in grams of CO2 per trip

In graph 16 the distribution of CO2 released per trip can be observed. The bicycle emission is compared to the car emission for the same distances. It is very predictable that vegan bike commuter has on average lower impact on CO2 emission than for instance regular commuter, an even bigger impact when compared to paleo commuter or car.

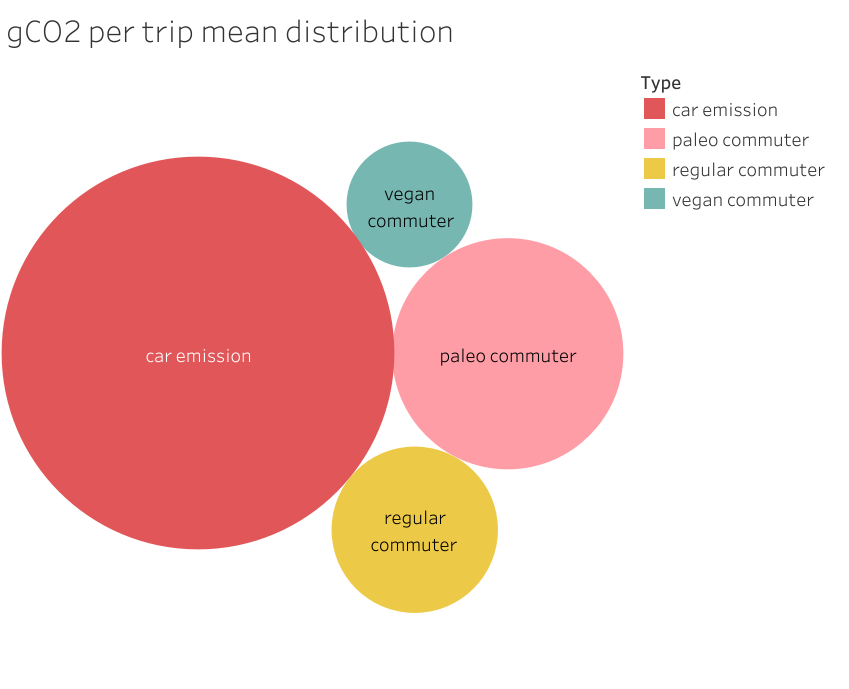
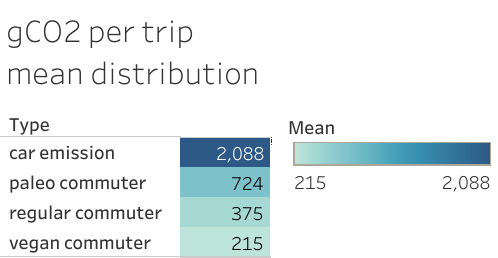


Figure 17 grams of CO2 distributions of mean per three different diets and car emissions



Rough estimates of biking impact show that following a vegan diet will lead to the most energy-efficient commuter (mean: 2015 gCO2 per trip). However, for instance, if a commuter is following a meat-heavy diet (mean: 724 gCO2 per trip), it will be the same level of efficiency to do a car share with two other individuals, or even reduces its impact by sharing a car with three (724gCO2 x 4 = 2896 gCO2/trip for the mean) or more people who are following the same diet.

Diet is important; however, the main concern lies in whether the overall usage of city bike share can affect climate change positively or not. For that, the average user count was calculated for each city.

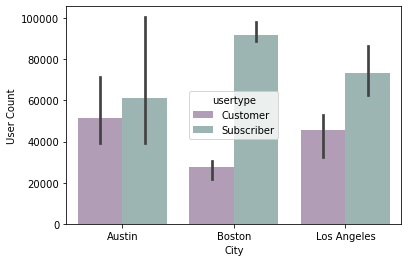


Figure 18 Number of Users and Subscribers per city

It can be observed on the graph above (Figure 18) that the number of subscribers is higher than day customers for every city with Austin having an almost equivalent distribution. It can be due to the fact that Austin has been seeing a lot of growth during those years (2019-2021), and remains mainly a car city due to the limited availability of bicycle routes. On the other hand, Boston (689,326) has almost 1/6th of Los Angeles (3.973 million) population but the number of users is significantly higher. Boston is primarily a walking city while Los Angeles is a car city. This information is important to consider, however, it does not play a valuable role in the analysis as the overall increase is the main target variable.

The total customer and subscriber user base was calculated for each year. Stable growth was observed between the years 2019 and 2021, however, there was a decline in usage due to the global pandemic ( -5.4% for customers and -8.2% for subscribers) during 2020. (Figure 19)

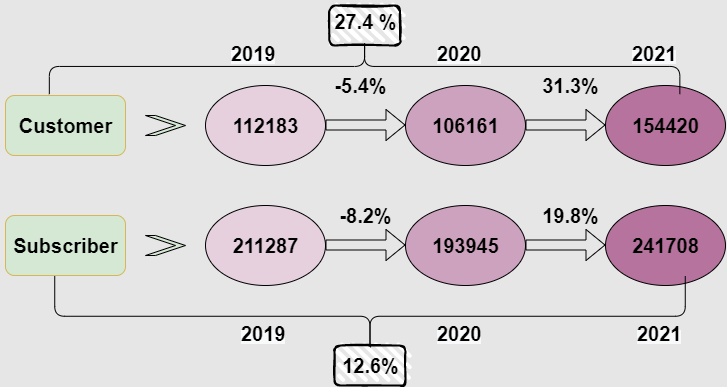


Figure 19 Flow chart of the percentage of bike usage between three years (2019-2021)

This leads to the introduction of the ARIMA (Autoregressive Integrated Moving Average) model. As mentioned in the introduction, this model was used to forecast time series for future city bike usage.

As the first step rolling mean and standard deviation was calculated (Figure 20). STD is a green (lower) line and rolling mean red (upper) line on the graph.

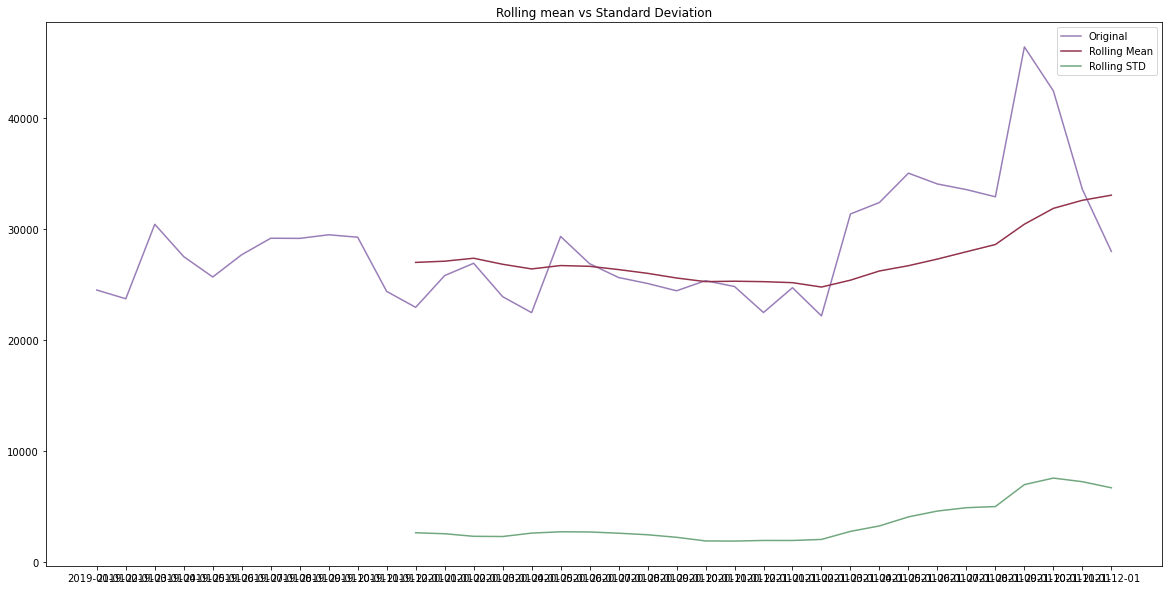


Figure 20 Rolling Mean vs Standard Deviation of the main data user count per month and year

The ideal data would have a rolling mean and standard deviation parallel to the x-axis. On the graph above it can be observed that they are indeed running parallel to the x-axis with a slight increase as time progresses. ADF statistics were performed next to analyze the results more accurately.

ADF Statistics: -2.4449231466628008

p-value: 0.12945638748215488

Critical value:

1%: -3.6327426647230316

5%: -2.9485102040816327

10%: -2.6130173469387756

From the statistics it can be concluded that the data is indeed stationary as ADF statistics is close enough to the critical values, however, the p-value is still not below the threshold and can be improved. The log of the dataset can be taken to render the data to transform to a more stationary form, alongside shifting the mean or subtracting the rolling mean, as potential they can improve the model.

Taking the log of customer count, subtracting the rolling mean, shifting the mean, and doing exponential decay all resulted in identical ADF statistics:

ADF Statistics: -2.4449231466628008

p-value: 0.12945638748215488

Critical value:

1%: -3.6327426647230316

5%: -2.9485102040816327

10%: -2.6130173469387756

Thus, the original data is in its most stationary form and will be used for continuous analysis.

Next ACF (Figure 21) for MA and PACF(Figure 22) for AR was calculated. As it was mentioned before, they both determine the number of terms used in the model.

Chart, histogram

Description automatically generated

Figure 21 ACF (Autocorrelation) model graph for the fitted dataset

The blue color identifies the horizontal threshold, therefore, only peaks above or below It are considered. For this dataset, there are only two peaks or terms that will be used for building the final ARIMA model.

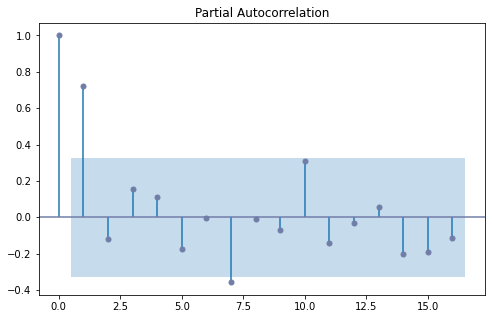


Figure 22 PACF (Partial Autocorrelation) for the fitted dataset

Here the PACF graph can be observed. Blue is a horizontal threshold, and only peaks outside of it are considered. For this model there 2 peaks, and thus, 2 terms used for future modeling. There is another peak, closer to the 7.5 value, however, it does not cross the threshold significantly enough to be considered for the model.

Thus, the fit for the ARIMA model is AR of order 2, differencing of order 1, and MA of order 2. The model then was run to predict fitted values.

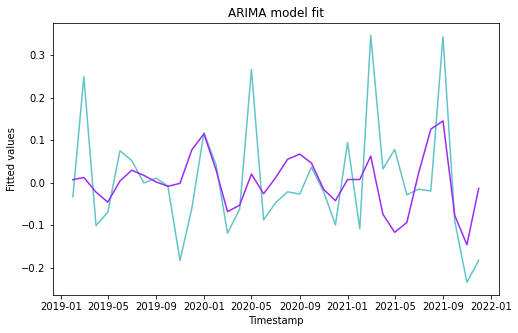


Figure 23 Fitted ARIMA model

On the graph above, the blue line shows the output of the original values, and the purple line the output of the predicted values to the same timestamps. The model accurately predicted the past outcomes, despite several sharp peaks. Those peaks could be explained by unusual activities that are outside of regular circumstances (World Cup, International Conference, COVID, etc.).

Figure 24 and Figure 25 show ARIMA predictions for 80 months ahead for the regular count and for the fitted count accordingly. The confidence interval shows that we could potentially expect up to 150000 users in mid-2023, however, the prediction is still that the usage will stay very consistent with regular seasonal drops and increases. There can be observed a slight overall growth in usage, however, it may not be significant enough to outweigh car CO2 emissions.

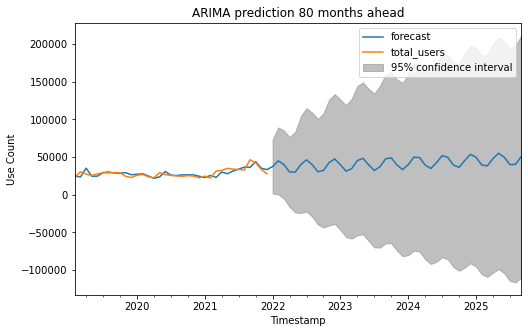


Figure 24 ARIMA predictive model for 80 months ahead using the original dataset

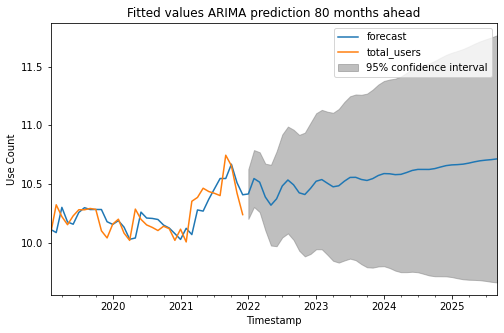


Figure 25 ARIMA predictive model for 80 months ahead using a fitted dataset

ARMA model was created for comparison. The ARMA model required the dataset to be split into the train (around 70%) and the test (30%). (Figure 26)

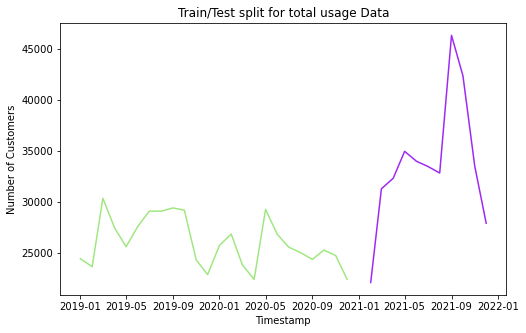
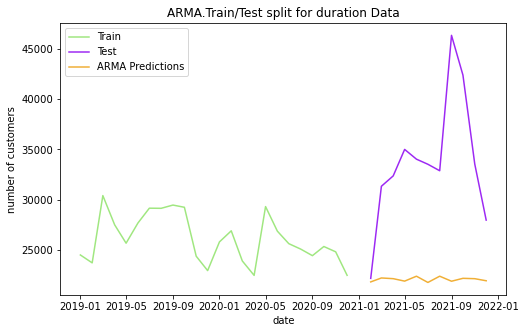


Figure 26 Train (green) and Test (purple) dataset distribution of the original data from the ARMA model

Chart, line chart

Description automatically generated

Figure 27 ARMA train and test split with the prediction of user number graph (left). Zoomed in prediction component (right)

On the graph above ARMA model with a predictive set is displayed. As it can be observed, the trend for prediction is linear, however, the RMSE for this particular model is equal to 13228.02. This number is very big to be considered a significant value. Therefore, this mode is not very well suited for this dataset.

**DISCUSSION**

The above analysis has demonstrated that city bike sharing programs have a positive impact on the carbon footprint, meaning it does help to reduce the amounts of gCO2 released in the atmosphere but not at significant amounts. The main determining factor in the effect of cycling on carbon emission is the number of users, as proportionally more people are using bicycles as their source of transportation, and fewer people are using cars and other modes of transit (all of them besides walking by foot emit more CO2 than bicycling – Figure 1). Three USA cities' data was used to construct the average usage and CO2 emission per different types of diets.

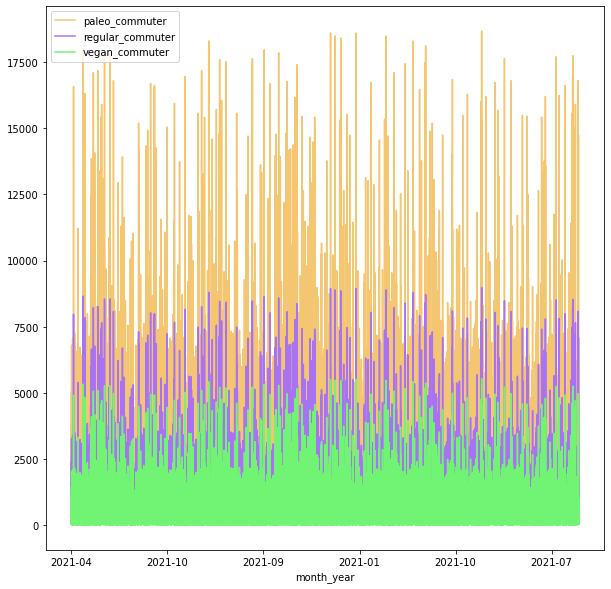


Figure 28 Random fractional data of gCO2 distribution per trip for three different diets

From the graph (Figure 28), it can be observed that a vegan diet has the lowest impact, and it can be suggested if individual cyclist wants to contribute positively to the sustainability of planet Earth, they should introduce more plant-based meals into their diet.

As mentioned above the number of users hypothetically plays the most significant role in adapting low carbon emission future. However, from the provided data, no significant growth in city bicycle usage was determined. That to say for bike sharing to have a high positive impact on emissions, bike sharing stations, and bike routes have to be more widely available and more affordable to attract a wider population of users. That is to say, government or private organizations perhaps need to subsidize it and promote using bike share as a main source of transportation. According to the National Household Travel Survey (NHTS), 90% of the American population is still using the car as a primary source of transportation, which is followed by walking – around 4.5% and only 0.6% of Americans are using bicycles as a mode to commute. 0.6 % or around two million people is not considered significant to have a reverse impact on carbon emission.

The average vehicle travels around 22 miles per gallon - to offset one gallon equal to the CO2 produced the average cyclist must bike around 430 miles. This is physically impossible for most human beings, especially considering the fact that the average commuter travels around 15 miles one way, and it will take them 14.5 trips to and from work to offset one gallon of gasoline. However, the intensity of CO2 emission during production has to be considered as well. The largest source of emissions during the production of both cars and bicycles are coming from steel. Bicycles, however, require much less steel (or aluminum), therefore, resulting in a small output of CO2 during the manufacturing process (*Figure 29*)

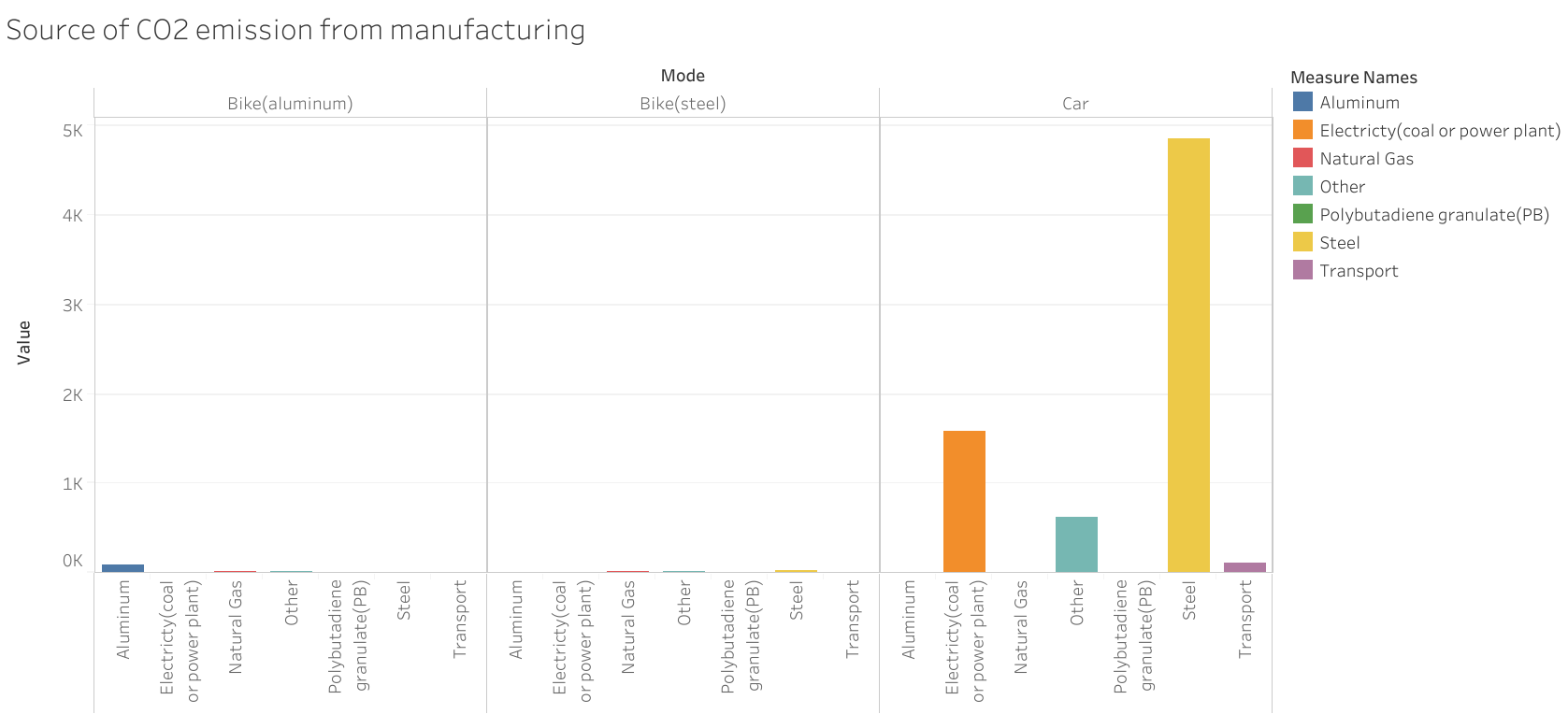


Figure 29 Sources of CO2 emissions for cars and bikes (aluminum and steel) during the manufacturing process in kgCO2 released

To conclude bicycles are still a great source of transportation and the benefits of using them outweigh using other modes, however, more than cycling has to be done just to reduce carbon emissions into the atmosphere. Not only does the number of users have to increase drastically, but people also should change their personal choices in diet to make bike commuting healthier for the environment than car driving. A big chunk (avoidable) of the carbon emissions comes from manufacturing and transportation of goods and perhaps the first step could be to reduce the usage of new things (whether it is a car or piece of clothes).

This study has shown that using city bikes does lessen the amounts of carbon released into the atmosphere. This project is undoubtedly valuable to the improvement of our world and future existence; however, more research has to be done. Only a small fraction of the points was taken into the account, and many more were neglected. Thus, it is only the first step (mostly exploratory) to a bigger problem and more data is required for a deeper and more meaningful analysis. Even though city bike usage is a good indication of the bike popularity, bike owners should also be taken into the consideration. In the ideal scenario, it would be preferable to collect information on people who are using their bikes as a source to commute places as well as for pleasure. Another important factor to add in the future is the relative analysis of cities' infrastructures – how accessible and available are bike routes. As well as economic analysis – how income and access to resources affect one’s choices in choosing a car or bike as a commute. The question of whether bike sharing can bring a cleaner future and how to do is certainly valuable and requires more hands and minds to come up with functioning and widely spread answers and solutions.

**LIST OF FIGURES**

* Figure 1 Average Carbon Emissions by Transportation Type (grams per kg) (page 2)
* Figure 2 Estimated Number of Bike Sharing Programs Around the World (c.2022) (page 3)
* Figure 3 Example of original data for single month entry and single city (New York) (page 6)
* Figure 4 Distance Formula (page 7)
* Figure 5 World Greenhouse Gasses Emissions in percentage (page 8)
* Figure 6 Stationary and Non-Stationary Time Series Example Graphs (page 9)
* Figure 7 Stationary and Non-Stationary Variance example graphs (page 10)
* Figure 8 Reduced Dataset Head Example (page 10)
* Figure 9 Total User Count Stationary Data with a Trend Component. The Trend Component is Not Distributed Evenly (page 11)
* Figure 10 Unit Root Test Formula (page 11)
* Figure 11 Augmented Dickey-Fuller Test Formula (page 12)
* Figure 12 ARMA Model Formula (page 12)
* Figure 13 Example of Data with Trend Component (left) and Differencing Component (right) (page 13)
* Figure 14 ACF (top) and PACF (bottom) graphs examples (page 14)
* Figure 15 ARMA (left) and ARIMA (right) Graphs for Comparison (page 15)
* Figure 16 Three different diets (regular, plant-based (vegan), meat-based (paleo)) and their relative CO2 emissions in grams of CO2 per trip (page 16)
* Figure 17 grams of CO2 distributions of mean per three different diets and car emissions (page 17)
* Figure 18 Number of Users and Subscribers per each city (page 18)
* Figure 19 Flow chart of the percentage of bike usage between three years (2019-2021) (page 19)
* Figure 20 Rolling Mean vs Standard Deviation of the main data user count per month and year (page 20)
* Figure 21 ACF (Autocorrelation) model graph for the fitted dataset (page 21)
* Figure 22 PACF (Partial Autocorrelation) for the fitted dataset (page 22)
* Figure 23 Fitted ARIMA model (page 23)
* Figure 24 ARIMA predictive model for 80 months ahead using the original dataset (page 24)
* Figure 25 ARIMA predictive model for 80 months ahead using the fitted dataset (page 25)
* Figure 26 Train (green) and Test (purple) dataset distribution of the original data from the ARMA model (page 25)
* Figure 27 ARMA train and test split with the prediction of user number graph (left). Zoomed in prediction component (right) (page 26)
* Figure 28 Random fractional data of gCO2 distribution per trip for three different diets (page 27)
* Figure 29 Sources of CO2 emissions for cars and bikes (aluminum and steel) during the manufacturing process in kgCO2 released (page 28)

**REFERENCES**

\*Data collection sources:

* “NYC DOT - Data Feeds.” *NYC*, 2022, www1.nyc.gov/html/dot/html/about/datafeeds.shtml#realtime.
* Chicago: “Divvy Trips | City of Chicago | Data Portal.” *Chicago*, 19 May 2020, data.cityofchicago.org/Transportation/Divvy-Trips/fg6s-gzvg.
* Austin: “Open Data | City of Austin Texas.” *Austin*, 2022, data.austintexas.gov/Transportation-and-Mobility/Austin-MetroBike-Trips.
* Los Angeles: “Data.” *Metro Bike Share*, 23 June 2022, bikeshare.metro.net/about/data.
* Boston: “Bluebikes System Data.” *Blue Bikes Boston*, 2022, [www.bluebikes.com/system-data](http://www.bluebikes.com/system-data).

“CDC Museum COVID-19 Timeline.” *Centers for Disease Control and Prevention*, 5 Jan. 2022, [www.cdc.gov/museum/timeline/covid19.html](http://www.cdc.gov/museum/timeline/covid19.html).

Sustainabilityinfo. “What Is the Average Speed for Urban Cycling?” *Urban Cycling*, 16 Feb. 2022, sustainabilityinfo.com/transportation/what-is-the-average-speed-for-urban-cycling.

Geus, De. “Determining the Intensity and Energy Expenditure during Commuter Cycling.” *British Journal of Sports Medicine*, 1 Jan. 2007, bjsm.bmj.com/content/41/1/8.full.

“Alternative Fuels Data Center: Maps and Data - Average Fuel Economy by Major Vehicle Category.” *U.S Department of Energy*, 20 Feb. 2020, afdc.energy.gov/data/10310.

Nealon, Lizzie. “Average Car Weight.” *Bankrate*, 4 Jan. 2022, www.bankrate.com/insurance/car/average-car-weight/#:%7E:text=The%20average%20weight%20of%20a,Environmental%20Protection%20Agency%20(EPA).

Warehouse, Bicycle. “How Much Does A Bicycle Weigh?” *Bicycle Warehouse*, 23 Feb. 2021, bicyclewarehouse.com/blogs/news/how-much-does-a-bicycle-weigh.

“Greenhouse Gas Emissions from a Typical Passenger Vehicle.” *US EPA*, 30 June 2022, www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle.

Scarborough, Peter. “Dietary Greenhouse Gas Emissions of Meat-Eaters, Fish-Eaters, Vegetarians and Vegans in the UK.” *SpringerLink*, 11 June 2014, link.springer.com/article/10.1007/s10584-014-1169-1?error=cookies\_not\_supported&code=6ee13f4f-0441-4469-a4a1-6a035be437b0.

Brownlee, Jason. “How to Remove Trends and Seasonality with a Difference Transform in Python.” *Machine Learning Mastery*, 22 June 2020, machinelearningmastery.com/remove-trends-seasonality-difference-transform-python.

“Introduction to ARIMA Models.” *Duke University*, Duke University, 18 Aug. 2020, people.duke.edu/%7Ernau/411arim.htm.

“The Environmental Impact of Today's Transport Types.” *TNMT*, 12 May 2021, tnmt.com/infographics/carbon-emissions-by-transport-type.

“National Household Travel Survey.” *National Household Travel Survey*, nhts.ornl.gov. Accessed 23 July 2022.

U.S. Census Bureau. “May 17 Is National Bike to Work Day.” *Census.Gov*, 8 Oct. 2021, www.census.gov/library/stories/2019/05/younger-workers-in-cities-more-likely-to-bike-to-work.html#:%7E:text=The%20most%20recent%20American%20Community,method%20of%20transportation%20to%20work.

Corporation, Trek Bicycle. “How Far Do You Need To Ride For Your Bike To Be Carbon Neutral? Trek Bicycle Reveals the Answer, Celebrating the Second Year of #GoByBike.” *Cision*, 20 May 2021, www.prnewswire.com/news-releases/how-far-do-you-need-to-ride-for-your-bike-to-be-carbon-neutral-trek-bicycle-reveals-the-answer-celebrating-the-second-year-of-gobybike-301296132.html#:%7E:text=The%20average%20vehicle%20travels%2022,their%20Trek%20bike%20430%20miles.