MathWorks Math Modeling Challenge 2023

High Technology High School

Team #16559, Lincroft, New Jersey

Coach: Raymond Eng

Students: Michael Gao, Kevin Guan, Amanda Guan, Amanda Lin, Kevin Liu

M3 Challenge THIRD PLACE—$10,000 Team Prize

JUDGE COMMENTS

Specifically for Team # 16559 —Submitted at the Close of Triage Judging:

COMMENT 1: Very nice paper. Your executive summary was well done and included results, as well as your detailed set of assumptions with justifications. Nice use of sensitivity analysis -- you didn't just say what you modified but its impact on the results. I like how you broke up and analyzed the UK and the US separately -- perhaps one would be a predictor of the other.

COMMENT 2: This was an excellent contribution! The executive summary was informative and well written. The response to the first and third portions were particularly impressive. Similarly, the response to the second component represented strong critical and analytical thinking skills. The sensitivity analysis and strengths and weaknesses subsections were also nice portions of the project, and the conclusion was well-formed. Fantastic job!

COMMENT 3: Use of a Bass diffusion model sounds interesting, but is not well explained and I was hoping for some more justification. It looks like a logistic model, based on the Figures. Among the predictors, surely imitation and innovation are not the only relevant features - what about the costs of the technology and other economic/feasibility measures? Overall, this is a highly creative proposal.

COMMENT 4: Your use of Bass Diffusion Model seems interesting. Team predicted that electric bike sales in the US will be 1.57 million in 2025 and 2.223 million in 2028, while sales in the UK will be 479 thousand in 2025 and 260 thousand in 2028. The factors guiding the e-bike sales being urban population and electricity prices. You also quantified the CO2 emission, time savings and health benefits. Looks like appropriate assumptions and justifications are made for the model. Graphs shown in fig 2.2.1 fits beautifully with the data. Glad to see the evaluations and calculations provided. Thanks for the sensitivity analysis and the showing the strength and weaknesses of the models. Excellent work.

Ride Like the Wind Without Getting Winded: The Growth of E-bike Use

Executive Summary

The popularity of electric bikes has been growing rapidly in recent years. E-bikes have started to become an attractive alternative to cars or public transit, and they have the potential to play a role in sustainable energy plans for the US Department of Transportation and UK Department of Transport. This paper proposes mathematically founded insights on the future growth of e-bike sales, factors influencing e-bike popularity, and the impact of increased e-bike usage—which can be used to advise the head of these transportation departments on policy decisions.

One of our goals is to forecast sales of new e-bike technologies in the United States and United Kingdom for the next two and five years using the Bass Diffusion Model. As this is a relatively new product, accurate predictions can be made by analyzing data on adoption rates. The model uses non-linear least squares regression to fit past data to the Bass Diffusion Model. We found that the coefficients of innovation and imitation in the US were 0.00237 and 0.2257 respectively, and in the UK, they were 0.00471 and 0.2775 respectively. By calculating the difference in the installed base fraction multiplied by the market cap over consecutive years, we predict that electric bike sales in the US will be 1.57 million in 2025 and 2.223 million in 2028, while sales in the UK will be 479 thousand in 2025 and 260 thousand in 2028.

The influence of various factors on people's adoption of e-bikes was explored using the random forest algorithm. In the results, urban population and electricity prices emerged as the most influential variables in driving the adoption of e-bikes in both the US and UK. Interestingly, the impact of disposable income varied between the two countries under consideration, with a significant effect observed in the United States, but not in the UK. On the other hand, the perception of the environment was found to have very little impact on e-bike adoption in either country. These findings offer valuable insights into the key drivers of e-bike adoption and could be useful for policymakers and businesses looking to promote sustainable transportation options.

The shift towards e-bike usage as a primary mode of transportation will have significant and long-lasting impacts on carbon emissions, traffic congestion, and public health in the United States and the United Kingdom. We have estimated the likelihood of commuters switching from driving cars, using public transportation, walking, or traditional cycling to e-biking based on convenience and cost. Using a formula to measure carbon emission savings, we predict that the US could save approximately $105,336,765.5$ metric tons of carbon emissions per year, while the UK could save $19,311,022.86$ metric tons. Because there will be less congestion, an average commuter in the US can save 4.425 minutes, and an average commuter in the UK can save 4.982 minutes. E-biking also has considerable health benefits for commuters, with an average of 482 more calories burned per day for US commuters and 369 more calories burned per day for UK commuters.

Electric bikes have the potential to revolutionize transportation in the United States and United Kingdom. This increasingly popular technology can help to reduce carbon emissions, alleviate traffic congestion, and improve overall well-being for commuters. By understanding the key factors that influence the growth of e-bikes, governments can strategically target, prioritize, and optimize areas to take full advantage of this novel technology. Our models provide invaluable insight into the exciting growth of e-bikes, helping to shape the future of our transportation in a positive and sustainable way.

Contents

1 Introduction

As e-bikes have been growing in popularity in recent years, we are tasked with modeling the projected growth, determining the significance of factors that impact said growth, and quantifying the impacts of the reduced usage of other modes of transportation in the United States and United Kingdom.

1.1 Restatement of the Problem

The problem we are tasked with addressing is as follows:

1. Create a mathematical model to predicts growth of e-bike sales two and five years from now in the US and UK.

2. Use mathematical modeling to determine the significance of several factors in the growth of e-bike usage in the US and UK.

3. Develop a model to quantify the impacts of reduced usage of other modes of transportation due to the increase of e-bikes in the US and UK.

1.2 Global Assumptions

1. *Consumers are rational beings.* This is a necessary assumption for us to model the behavior of consumers mathematically.

2. *There are no new technological advancements or government policies that significantly impact the number of e-bike users over the next five years.* New technological advancements and new government policies are unpredictable and thus difficult to predict. Therefore, it is necessary to assume the number of bike users stay constant to simplify our models.

3. *The total population will remain constant.* While there will be population and age demographic changes, these changes will be negligible in the short-term, allowing for this simplifying assumption.

2 Part I: The Road Ahead

E-bikes have been greatly increasing in popularity as they have become an attractive alternative to other forms of transportation. In this section, we create a mathematical model to predict the growth of e-bike sales two and five years from now in the US and UK.

2.1 Assumptions

1. *No one who already bought an e-bike will need to buy a new e-bike.* Accounting for e-bike users replacing their vehicles is beyond the scope of the model. This assumption is reasonable given that we are only predicting sales in the next five years.

2. *Consumers that do not use a bike will not buy the e-bike technology over the next five years.* While it is possible that someone who does not already own a bike will want to buy an e-bike bike, we limit the scope of our model to those who currently regularly use bikes for transportation, as this is the primary target demographic of e-bikes.

3. *Information about electric bikes spreads via word-of-mouth and advertising. The rate of e-bike adoption based on these methods is constant over the next five years.* There is no reason to believe that the effect of advertising and word-of-mouth on e-bike adoption will change in the next five years.

4. *All people who regularly use bicycles as a mode of transportation will eventually buy an e-bike.* It is reasonable to assume that only people who bike for transportation, as opposed to recreation, will invest in an e-bike. Given pressures to migrate to greener technology and the increasing cost efficiency of electric transportation, this assumption is necessary to simplify the model.

5. *The proportion of bicycle users who use e-bikes in the UK is equal to the proportion of bicycle users who use e-bikes in Europe.* Data on e-bikes sold is available for Europe but not the UK. We assume the proportion of e-bike users in Europe is representative of the UK.

6. *A person buying an e-bike is equivalent to that person adopting e-bikes and replacing bicycles as their new mode of transportation.* This is consistent with assumption assumption 1, as it implies that there are no e-bike purchases from existing e-bike users. Defining the adoption of e-bikes this way is necessary for the application of our model.

2.2 Model Development

To predict the adoption of new technologies such as the e-bike, we use the Bass Diffusion model, which categorizes consumers of the population as "innovators" and "imitators." [1] The formula for the model is given below:

$$F(t) = \frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p} e^{-(p+q)t}}.$$

with the following definitions:

- $F(t)$ represents the proportion of the market using the product,
- p is the coefficient of innovation, i.e., the rate at which the market is adopting the product via advertising,
- and q is the coefficient of imitation, i.e., the rate at which the market is the adopting the product via word-of-mouth from current users.

Because the coefficients of innovation and imitation are likely different between the US and U.K, we perform two separate regression analyses for each location.

First, we found data for the total number of people who regularly used bikes for transportation in each location, which we take as the potential market size because of assumption 5. To find this value for all of Europe, we multiplied the percentage of regular bike users by the population in each European country for which data was available and added them together. This calculation is shown n in the table below for the five countries with greatest proportions of bike users. For the UK and the US, we were able to directly find these values.

Table 2.2.1: Bike Users for European Countries

Country	Bike User Proportion	Population	Number of Bike Users
Netherlands	0.36	17590672	6332642
Denmark	0.23	5873420	1350887
Hungary	0.22	9689010	2131582
Sweden	0.17	10452436	1776914
Finland	0.14	5548241	776754

Table 2.2.2: Bike Users in Target Locations

Location	Total Bike Users (thousands of users)
United States	45000 [2]
United Kingdom	10700 [3]
Europe	40794

To develop our model, we used the number of annual e-bike sales in the US and Europe [4]. For the U.S, data for 2012 to 2017 was found [5] and used in addition to the provided data from 2018 to 2022. This allows us to calculate the proportion of bicycle users who have adopted e-bikes with the following equation:

$$F_t = \sum_{0}^{t} \frac{s_t}{m}$$

where:

- F is the proportion of bike users who have adopted e-bikes,
- t is the the number of years since the first year of data collection,
- s is the number of bike sales,
- and m is the total number of regular bicycle users.

The first five years of these calculations are displayed in the table below for the U.S and Europe.

Table 2.2.3: E-bike Users in the U.S for 2012-2016

Year	Sales (thousands)	Change in E-bike Proportion	Total E-Bike Proportion
2012	70	0.001556	0.001556
2013	159	0.003533	0.005089
2014	193	0.004289	0.009378
2015	130	0.002889	0.012267
2016	152	0.003378	0.015644

Table 2.2.4: E-bike Users in Europe for 2006-2010

Year	Sales (thousands)	Change in E-Bike Proportion	Total E-Bike Proportion
2006	98	0.002402	0.002402
2007	173	0.004241	0.006643
2008	279	0.006839	0.013482
2009	422	0.010345	0.023827
2010	588	0.014414	0.038241

This completes the analysis necessary to develop a data set that we can use for our model. Fitting the equation for the Bass diffusion model using a non-linear least squares regression to this data yielded the following values for p and q in each location:

Table 2.2.5: Estimated Parameters of the Bass Diffusion Model

Location	Coefficient of Innovation (p)	Coefficient of Imitation (q)
United States	0.00237	0.2257
Europe	0.00471	0.2775

The below figures show the Bass diffusion models graphed along side historical data and extrapolated to 2029. The points at two (2025) and five (2028) years from now are labeled.

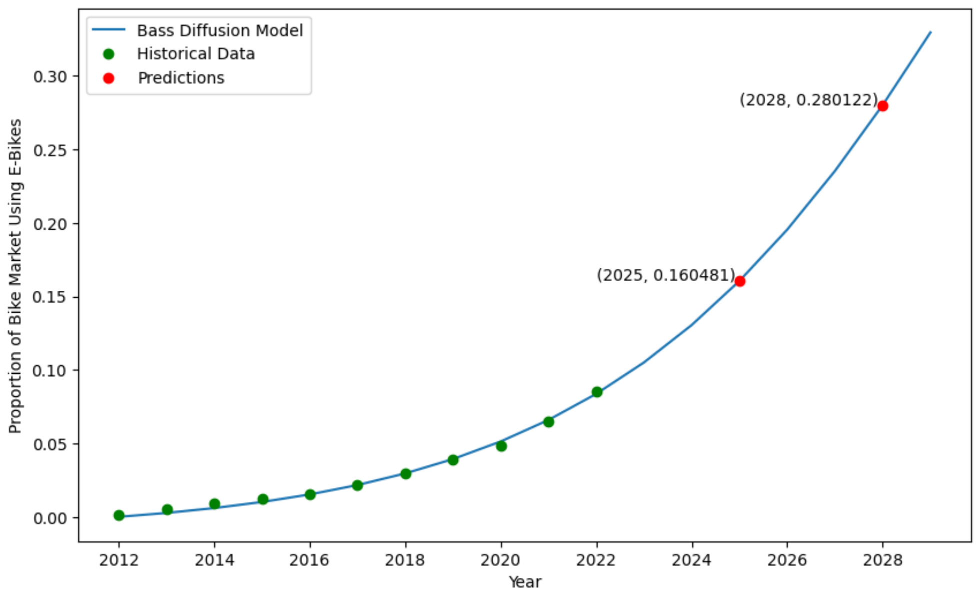

Figure 2.2.1: Proportion of US Bicycle Market Using E-Bikes

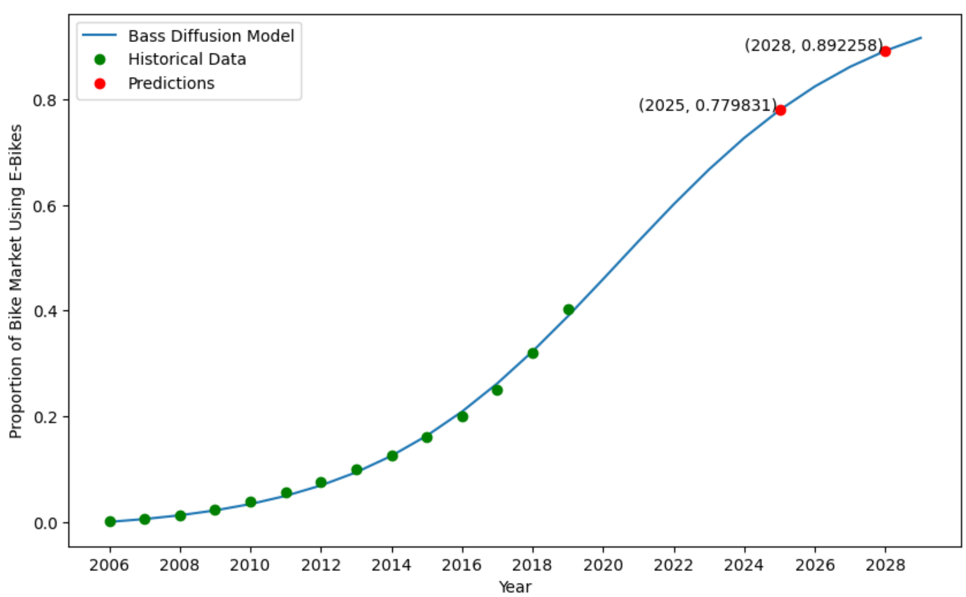

Figure 2.2.1: Proportion of European Bicycle Market Using E-Bikes

2.3 Results

The values produced by this model can then be used to calculate the number of bikes sold using the following formula:

$$s_t = m[F(t+1) - F(t)]$$

- s is the number of bike sales,
- t is the number of years since the first year of data collection
- $F(t)$ is the proportion of bike users who have adopted e-bikes according to the Bass diffusion model,
- and m is the total number of regular bicycle users.

This formula essentially reverses the conversion of the original data performed in Table 2.2.3 and Table 2.2.4. The model was developed using the data for Europe because of assumption 5, so we now use the number of bike users in the UK from table 2.2.2 for m. This calculation is shown in the table below using the values for 2025-2026 and 2028-2029.

Table 2.3.1: Predictions of Number of E-Bike Sales in 2025 and 2028

Country	Year t	Year t+1	F(t)	F(t+1)	*F(t+1)-F(t)*	E-Bike Sales (thousands)
US	2025	2026	0.168481	0.195359	0.030004	1570
US	2028	2029	0.280122	0.329520	0.044853	2223
UK	2025	2026	0.779831	0.824632	0.044802	479
UK	2028	2029	0.892258	0.916577	0.024319	260

2.4 Sensitivity Analysis

To analyze the sensitivity model, we adjust the values of p and q individually by 10%. The results of the analysis are shown below.

Table 2.4.1: Results of Sensitivity Analysis with Constants p and q

Country	Constant	Change (%)	Change in 2025 Sales (%)	Change in 2025 Sales (%)
U.S.	p	+10%	+1.1486%	+0.5688%
U.S.	p	-10%	-1.2386%	-0.6693%
U.S.	q	+10%	+4.3439%	+2.9642%
U.S.	q	-10%	-3.7235%	-2.9377%
U.K.	p	+10%	-0.3276%	-0.2128%
U.K.	p	-10%	+0.3623%	+0.2488%
U.K.	q	+10%	-1.0213%	-0.7976%
U.K.	q	-10%	+0.7696%	+0.9044%

In varying p and q by 10%, e-bike sales do not change much in either 2025 or 2028. We can see that e-bike sales respond more to changes in q than changes in p, which makes sense given the meanings of p and q. Technological adaption caused by word-of-mouth increases with the number of users, while adaption as a result of advertising does not.

The directions of for p and q corresponded to those of e-bike sales in the U.S, which was expected, as larger coefficients of technological adaptation should increase the change in users. We were initially surprised to see that, in the U.K, the directions of changes in p and q were opposite the resulting changes in e-bike sales. However, in examining Figure 2.2.1, we can see that the proportion of e-bike users has already crossed an inflection point and is concave downward as it approaches the market size, which acts as the model's carrying capacity. The results are therefore logical because larger coefficients of technological adaption should cause the proportion of e-bike users in the U.K. to approach the market size more quickly.

2.5 Evaluation and Verification

The average values of p and q are 0.03 and 0.38 respectively; p is often 0.01 or less, while q is typically between 0.3 and 0.5. Considering these values, the estimates for the parameters p and q in table 2.2.5. In both the US (0.00237) and UK (0.00471), p is less than 0.01 and differs from the average value 0.03 by only one order of magnitude. The values of q (0.2257 in the US and 0.2775 in the UK), while not between the typical bounds of 0.3 and 0.5, are still not drastically different from the average value of 0.38. The fact that they are relatively low is also reasonable given that there are many reasons for consumers to be hesitant to switch to electric modes of transportation [6].

2.6 Strengths and Weaknesses

The Bass Diffusion model has many strengths in predicting the adoption of new technologies. With the coefficients of innovation and imitation, Bass Diffusion models information spread via word-of-mouth and advertising, incorporating both external and internal growth factors in the model. In addition, the Bass Diffusion model can be scaled up or down to fit different market sizes, such as the US and UK markets for electric bicycles.

One weakness of the model is that it assumes the all regular bike users will eventually adopt the new technology, which is not necessarily the case in the real world. It also does not account for new e-bike users who were not initially bike users. Another weakness of our Bass Diffusion model is that it does not extrapolate well in the long-term. The number of bike users, or "market cap," was assumed to stay constant over the next five years, but with rising populations and bike production, this would not be the case many years down the line. Finally, the Bass diffusion itself has a limited scope. It assumes that the only two factors in the adoption process are innovation and imitation, failing to take into consideration other influences such as social norms and regulatory barriers.

3 Part II: Shifting Gears

There are many factors that influence people's choice to switch to e-bikes and therefore lead to the growth of e-bike usage. These factors include gas and electricity prices, environmental awareness, commute times, and more. In this section, we model the the significance of several factors in explaining the growth of e-bike usage in the US and UK.

3.1 Assumptions

1. *US State policies offering incentive for buying e-bikes are insignificant compared to national policies in the US.* This is a simplifying assumption that is reasonable because state policies are nonuniform and often have narrow reach or specific criteria. National policies would have greater effect.

2. *There are no government policies in the UK that significantly impact e-bike usage.* After extensive research, we were unable to determine any widespread UK government policies that would affect all people equally.

3. *The US inflation rate is equal to the UK inflation rate.* This is a simplifying assumption, as these two inflation rates are relatively close to each other, and allows for one inflation rate (the US rate) to be used across all calculations. [7].

4. *The growth in bike usage can be measured by the growth in bike sales.* This is a simplifying assumption based on the data we have available.

3.2 Model Development

In developing our model, we determine key factors that influence the growth of e-bike usage and use a random forest algorithm to order their significance by feature importance.

3.2.1 Factor Identification

First, we identify a variety of factors that are likely to influence the growth of e-bike usage. Our chosen factors that we used in our model are listed below.

1. *Gas prices:* High gas prices would encourage consumers to switch to e-bikes to save money.

2. *Electricity prices:* High electricity prices would discourage consumers from switching to e-bikes, while low electricity prices would encourage them to switch.

3. *Disposable income:* If a consumer has a greater amount of disposable income, they may be more willing to make an investment into e-bikes.

4. *Government incentives:* Incentives such as rebates for owning e-bikes would encourage consumers to buy and use e-bikes.

5. *Environmental perceptions:* If consumers care about the environment, they would be more likely to use e-bikes as an environmentally-friendly alternative.

6. *Urban population:* Consumers are more likely to use e-bikes in urban environments because travel distances tend to be shorter and there is typically better biking infrastructure.

7. *Number of bikeshare systems:* A greater number of bikeshare systems would encourage more people to use e-bikes from these systems. An increasing number of these systems is also likely to reflect a trend of increasing e-bike popularity.

3.2.2 Collecting Input Data

Table 3.2.2 lists each of the input factors and how they are defined.

Table 3.2.1: Representation for Each US Input Factor

Factor	Representation
Gas Prices	USD per gallon adjusted for inflation [4]
Electricity Prices	USD per kWh [4]
Disposable Income	Chained 2012 USD [4]
Government Incentives	Yes (1) or No (0) [10]
Environmental Perceptions	% who care a ”great deal” about the environment [4]
Urban Population	% of total population living in urban setting
Number of Bikeshare Systems	Number of docked and dockless bikeshare stations

Table 3.2.2: Representation for Each UK Input Factor

Factor	Representation
Gas Prices	Pence per liter adjusted for inflation [4]
Electricity Prices	Change in cost based on the consumer price index [4]
Disposable Income	Chained 2021 GBP
Environmental Perceptions [4]	% who included the environment as a top 3 important issue
Urban Population	% of total population living in urban setting
Number of Bikeshare Systems	Number of Bikeshare Bike Hires in London

3.3 Results

After training random forest models on the UK and US data separately, we used the feature importance attribute to find the importance of each factor. We ranked then ranked each factor by these calculated importances for each location. It is important to note that we focus on the order of importance of these factors. Given more time, we could explore the mathematical significance of the feature importances and how they are calculated. The results of our analysis are displayed in the tables below.

Table 3.3.1: Importance of Each Factor for the US

Factor	Rank	Importance
Disposable Income	1	0.249776
Urban Population	2	0.228853
Electricity Prices	3	0.199010
Number of Bikeshare Systems	4	0.192196
Government Incentives	5	0.057273
Gas Prices	6	0.0442675
Environmental Perceptions	7	0.028625

Table 3.3.2: Importance of Each Factor for the UK

Factor	Rank	Importance
Electricity Prices	1	0.244375
Urban Population	2	0.240936
Gas Prices	3	0.185762
Number of Bikeshare Systems	4	0.175073
Environmental Perceptions	5	0.088524
Disposable Income	6	0.065330

3.4 Sensitivity Analysis

To analyze the sensitivity of our model, we dropped each input feature one at a time and retrained our model with the remaining features only. We ranked the factors by their feature importances and subtracted the new ranks from the old ranks. These results are shown in Tables 3.4.1 and 3.4.2.

All of the changes were minimal, with magnitudes generally around 0-2 ranks and with only one change of magnitude 3 in disposable income in the US. Many of changes can also be explained by the dropping of one of the columns. For example, the disposable income in the UK had to increase in rank by 1 each time because it was in the last rank previously, so each time a column is dropped it would automatically increase by 1 in rank.

The minimal changes in rank suggests that our model is robust against changes, rendering validity to our model results.

Table 3.4.1: Change in Rank of Each Factor for the US

	Number of Bikeshare Systems	Electricity Prices	Urban Population	Gas Prices	Environmental Perceptions	Disposable Income	Government Incentives
Rank	—	0	1	1	1	-1	1
	1	—	1	1	1	-1	1
	2	2	—	0	2	-2	1
	1	2	-2	—	1	-1	0
	2	2	-1	0	—	-3	0
	1	2	0	1	1	—	1
	0	1	-1	0	2	0	—

Table 3.4.2: Change in Rank of Each Factor for the UK

	Number of Bikeshare Systems	Electricity Prices	Urban Population	Gas Prices	Environmental Perceptions	Disposable Income
Rank	—	-1	1	0	1	1
	2	—	1	0	1	1
	2	0	—	0	1	1
	1	-1	1	—	1	1
	1	-1	1	-1	—	1
	1	-1	1	-1	0	—

3.5 Evaluation and Verification

The mean absolute error (MAE) of the random forest algorithm averages the distance (i.e., error) between the actual and predicted data points.

$$\Sigma_{i=1}^{n} \frac{|Error|_i}{n}$$

Table 3.5.X: Mean Absolute Error

Country	MAE
US	21.736
UK	28.932

Theses MAEs are relatively small compared to the data values, which makes this model a good fit for the data.

3.6 Strengths and Weaknesses

The model has several strengths, including a low mean absolute error and the ability to effectively compare the relative influence of different factors on the decision to choose e-bikes. However, it also has notable weaknesses, such as the inability to evaluate the magnitude of each factor and the lack of transparency in the black box nature of the Random Forest algorithm.

4 Part III: Off the Chain

There are an increasing number of people who prefer e-bikes as their primary mode of transportation, which overall decreases other modes of transportation such as cars, buses, bikes, and walking. This section aims to quantify the impacts of switching to e-bikes on carbon emissions, traffic congestion, and health and wellness in the United States and United Kingdom.

4.1 Assumptions

1. *The average number of miles for each mode of transportation taken per person is the same for both the US and the UK.* This is a simplifying assumption due to a lack of data for the UK. Additionally, because of the Central Limit Theorem, over large populations, the average number of miles traveled for a given mode of transportation will tend towards the average, so to simplify the model, we assume that each person travels the average number of miles for their preferred mode of transportation.

2. *The fuel efficiency of each mode of transportation are constant in the US and UK.* Because of the Central Limit Theorem, the fuel efficiency will tend towards the average, so we can assume that the fuel efficiency is constant across that particular mode of transportation.

3. *The fuel used in each mode of transportation yields the same amount of carbon emissions per gallon.* On average, the same amount of fuel will release the same amount of carbon.

4. *The percent of people working from home in UK is equal to that in the US.* As there was no data provided for the percent of people working from home in UK, we assume that the percent of people would be equal, and adjusted the

5. *People who work from home do not do any traveling outside of their house, as the main reason for people to travel is to commute to their jobs.* This simplifies the model as we do not have data for people who travel outside of their commute.

6. *A consumer who switches from their preferred transportation method to e-bikes will travel the same number of miles per day.* This is because they will still be commuting to the same work location.

7. *When making decisions between transportation methods, consumers tend to want more convenience and lower cost.* This follows from our assumption that consumers are rational.

8. *The average cost of a vehicle is the same in the US and the UK.* This is a simplifying assumption that allows for a more consistent model across the two countries.

9. *Motorcyclists and rail riders do not switch to e-bikes.* E-bikes do not provide any convenience increase for these riders. For motorcyclists, their routes are much the same as an e-bike, and they would not want to purchase a new product. Meanwhile, rail riders travel a longer distance than would be reasonable for a person to travel on a e-bike.

10. *People who walk as their primary mode of transportation do not switch to e-bikes.* The cost of e-bikes is incredibly high compared the cost of walking, which is free, which makes it unfeasible to buy an e-bike.

11. *The speed drop due to congestion is constant for both the US and the UK.* This is a simplifying assumption due to a lack of data.

12. *There is no traffic no matter the number of bicycles on the road.* This is a simplifying assumption due to a lack of data.

4.2 Model Development

We analyze the effect of people switching from their preferred mode of transportation, namely car, bus/coach/public transportation, walking, bicycling, and working from home to electric bikes on three features:

- carbon emissions,
- traffic congestion,
- and health and wellness.

We use each transportation's average number of miles traveled per day in each of the three factors. We have the global constants below:

Table 4.2.1: Average Number of Miles Traveled Per Day Per Each Mode of Transportation

Vehicle	Average number of miles traveled per day per vehicle
Car	35
Public Transportation	20
Walk	3.5
Bicycle	15

Carbon Emissions: Consider the functions in the following table:

Table 4.2.2: Functions Used for Carbon Emissions

Function	Definition
M(T)	average number of miles traveled per day per person for each transportation T
C(T)	average carbon emissions of each transportation method per mile (gC per mile)
Pr(T)	probability of a user switching to an e-bike from T
Pe(T)	percentage of population that prefers transportation mode T
P(T)	average number of users per transportation mode T
F(T)	average carbon emissions saved per person

Notice that M(T)C(T) represents the total amount of carbon emitted per day for each vehicle. We have to divide this by P(T) to account for the carbon emitted per day per person. By subtracting away M(T)C(e-bike), we can get the emissions saved by a consumer who switched to e-bike transportation. However, we have to multiply by the probability Pr(T), as not every consumer is guaranteed to switch.

Thus, we have the following formula:

$$\mathrm{F}(T) = \left(\mathrm{M}(T)\frac{\mathrm{C}(T)}{\mathrm{P}(T)} - \mathrm{M}(T)\mathrm{C}(\text{e-bike})\right) \cdot \mathrm{Pr}(T).$$

Note that the carbon emissions per gallon of gasoline is 8887 grams of carbon. The average fuel efficiency of 24.2 miles per gallon for cars and 6.1 miles per gallon for buses/public transportation, so $\mathrm{C(car)} = \frac{8887}{24.2} = 367.2$ gC per gallon and $\mathrm{C(public\ transportation)} = \frac{8887}{6.1} = 1457$ gC per gallon. Therefore, from research and datasets collected, we have the following constants for vehicles in the US and UK:

Table 4.2.3: Transportation Constants in the US

Vehicle	Pe(T)	M(T)	C(T)
Car	75.1	35	367.2
Public Transportation	2.5	20	1457
Walk	2.2	3.5	0
Bicycle	0.4	15	0

Table 4.2.4: Transportation Constants in the UK

Vehicle	Pe(T)	M(T)	C(T)
Car	68.1	35	367.2
Public Transportation	6.2	20	1457
Walk	11.4	3.5	0
Bicycle	3.6	15	0

To calculate the probability to switch from transportation to e-bikes, we consider the "convenience" and cost factors. We measure convenience based on the speed (mph) of their preferred transportation method, where consumers tend to want higher speed and thus faster travel times. Per assumption 7, we assume consumers also tend to want a product with lower cost, and we weight both of these conditions equally.

Let $P(C)$ be the total population for each country C. Then $\text{Pe}(T) \cdot P(C)$ represents the total number of people that use T as their preferred mode of transportation. Then, we have $\text{Pe}(T) \cdot P(C) \cdot \text{F}(T)$ be the total amount of kilograms of carbon saved per day. Note the estimated commuter populations for each country below.

Table 4.2.5: Estimated Commuter Population of US and UK

Country	$P(C)$
US	155,284,955
UK	31,501,464

Traffic Congestion: We only analyze the conversion of people from cars to electric bicycles, as there is negligible congestion for people who walk, ride bikes, or take public transport.

Table 4.2.6: Functions and Variables for Traffic Congestion

Function	Definition
M(T)	average number of miles traveled per day per person for each transportation T
Pr(T)	probability of a user switching to an e-bike from T
Pe(T)	percentage of population that prefers transportation mode T
P(T)	average number of users per transportation mode T
S(T)	optimal speed for transportation mode T
P_c	proportion of the total population who commute
s_c	speed drop for a city
s_{max}	maximum speed driven in a city during the day (mph)
s_{min}	minimum speed driven in a city during peak traffic congestion hours (mph)
s_p	speed drop proportion per person

We find the weighted average by population of the speed drop of the 20 most populous US cities, as the number commuters in each city correspond with the population. The speed drop per city, s_c is defined as $s_c = \frac{s_{max}-s_{min}}{s_{max}}$, where s_{min} is the minimum speed during the peak traffic congestion hours, and s_{max} is the maximum speed driven throughout the day; without traffic congestion, a car would drive the maximum speed.

Table 4.2.7: Speed Drop by City due to Congestion (First Five Cities)

City	Speed Drop [8]	Population (2021) [9]
Boston	0.40	654,776
NYC	0.38	8,467,513
Miami	0.37	439,890
DC	0.35	670,050
San Fran	0.34	815,201

Using the weighted average speed drop per city, s_c, we find the speed drop proportion per person, s_p:

$$s_p = \frac{s_c \cdot P(\text{car})}{P(\text{country}) \cdot P_c \cdot Pe(\text{car})}.$$

Then, the change in commute time for the average commuter who would optimally be driving $M(\text{car})$ miles at $S(\text{car})$ miles per hour is computed. This is found by:

$$\Delta t = \frac{M(\text{car})}{S(\text{car}) \cdot s_p \cdot P(\text{car})} - \frac{M(\text{car})}{S(\text{car}) \cdot s_p \cdot P(\text{car}) \cdot Pr(\text{car})}.$$

Health and Wellness: For health and wellness, we consider the calories burned from riding an electric bicycle compared with the calories burned through other modes of transportation, in terms of the average miles traveled per person per day.

Table 4.2.8: Variable/Function Names and Definition

Function	**Definition**
P	Percentage of people using a particular mode of transportation
M	Miles traveled per day per person
CTransport	Calories burned through modes of transport other than e-bikes
CEbike	Calories burned through riding e-bikes
DeltaC	Difference between CEbike and CTransport
WCalories	Weighted DeltaC with respect to P percent of users
AvgCalories	Average difference in calories burned per person over all modes of transport

To calculate the number of calories burned from traveling M miles on an e-bike, we assume e-bikes travel at 20 miles per hour, and an average human burns 6 calories per minute on an e-bike [**flyer-bikes**]. Thus, the formula for the calories burned for e-bike riders is:

$$\text{CEbike} = \frac{M}{20} \cdot 6 \cdot 60 = 18M.$$

We then find the difference in the number of calories burned when switching to e-bikes, by subtracting the calories burned through each mode of transportation, CTransport, and CEbike:

$$\text{DeltaC} = \text{CEbike} - \text{CTransport}.$$

Now, consider P for ever mode of transportation. For the average person, we can take the weighted average of all of these modes of transportation to find AvgCalories.

$$\text{AvgCalories} = \sum \frac{\text{P}}{100} \cdot \text{DeltaC}.$$

Thus, substituting gives the formula of:

$$\text{AvgCalories} = \sum \frac{\text{P}}{100} \cdot (18M - \text{CTransport}).$$

For the United Kingdom, we proceed similarly. We assume that 17.9% of people work from home, and then assume that the car, bicycle, bus/coach, and walking sectors make up the other $100 - 17.9 = 82.1\%$ of the population. Thus, the new P can be calculated using:

$$P_{new} = (100 - 17.9) \cdot \frac{P}{\sum P}.$$

4.3 Results

We have the following probabilities for switching from preferred transportation, which will be used across multiple parts of the results:

$$\text{Pr(car)} = \frac{20}{20 + 45} \cdot \frac{40000}{2000 + 40000} = 0.4233$$

$$\text{Pr(bike)} = \frac{20}{12.5 + 20} \cdot \frac{525}{525 + 2000} = 0.1280$$

$$\text{Pr(public transportation)} = \frac{20}{12.7 + 20} \cdot \frac{3511.8}{3511.8 + 2000} = 0.3897$$

$$\text{Pr}(0) = \frac{20}{3.5 + 20} \cdot \frac{0}{0 + 2000} = 0.$$

Therefore, with C(e-bike) = 4.6, we can calculate the carbon emissions saved per person, on average, as

$$\begin{aligned} \text{F(car)} &= 2.4535 \text{ kgC per day per person} \\ \text{F(bike)} &= -0.010748 \text{ kgC per day per person} \\ \text{F(public transportation)} &= 0.14560 \text{ kgC per day per person} \\ \text{F(walk)} &= 0 \text{ kgC per day per person} \end{aligned}$$

Carbon Emissions: Summing over all the kilograms of carbon saved for each transportation method provided, we have that the average carbon emissions saved in the US is:

$$\sum_{\text{transportation} \in T} \text{Pe}(T) \cdot P(\text{US}) \cdot \text{F}(T) = 105{,}336{,}765.5 \text{ metric tons C per year},$$

which is equivalent to 22,899,297 cars' emissions saved per year for more context. Similarly, for the UK, we have

$$\sum_{\text{transportation} \in T} \text{Pe}(T) \cdot P(\text{UK}) \cdot \text{F}(T) = 17{,}751{,}419.19 \text{ metric tons C per year},$$

which is equivalent to $3,859,004$ cars' emissions saved per year for more context.

Traffic Congestion: The weighted average for speed drop of a city, s_c, was computed to be 0.3075. Based on the population for a country and the current percent of people who drive during their commute, we compute s_p:

Table 4.3.1: Speed Drop Proportion per Person

Country	s_p
US	$4.346110132 \cdot 10^{-8}$
UK	$4.824756677 \cdot 10^{-8}$

Then, using the calculated probabilities for switching from preferred mode of transportation to e-bikes, we compute the amount of time saved by an average commuter when that percent of people switch from a car to an e-bike.

Table 4.3.2: Commute Times for the US and UK (minutes)

Country	Original Commute Time	New Commute Time	Δt
US	53.904	49.479	4.425
UK	54.764	49.783	4.982

Therefore, an average commuter driving on the roads will save 4.425 and 4.982 minutes in the US and UK, respectively, from the shift to e-bikes.

Health and Wellness: For the United States, we can produce the following table:

Table 4.3.3: Calories Burned by Transportation in US

Transportation	**P**	**M**	**CTransport**	**WCalories**
Car	75.6	35	0	476
Public Transportation	2.5	20	0	9
Walking	2.2	3.5	350	-6.31
Biking	0.4	15	750	-1.92
Taxicab/Other	1.5	20	0	5.4
Work From Home	17.9	0	0	0

Taking the sum of the values in the WCalories column gives AvgCalories = **482** calories. For the United Kingdom, we calculate the rescaled P for each sector other than work from home to produce the following table:

Table 4.3.4: Calories Burned by Transportation in UK

Transportation	**P**	**M**	**CTransport**	**WCalories**
Car	62.6	35	0	394
Public Transportation	5.7	20	0	20.5
Walking	10.5	3.5	350	-30.1
Biking	3.3	15	750	-15.8
Work From Home	17.9	0	0	0

Taking the sum of the values in the WCalories column gives AvgCalories = **369** calories.

Thus, in total, the average person who switches to e-bikes in the US burns 482 more calories than they did with their prior mode of transportation. Similarly, the average gain in caloric burn in the United Kingdom is 369 calories.

4.4 Sensitivity Analysis

Carbon Emissions: To perturb the model, decrease the amount of car users by 5% arbitrarily, and increase the amount of other modes of transportation equally to make their total stay at 100%.

Scaling the US car commuter population down yields a car transportation percent of $75.6 \cdot 0.95 = 71.8$ percent, a 3.8% change. The rest of the sectors add up to $100 - 75.6 = 24.4\%$, implying a $3.8/24.4 = 0.156\%$ change every percent of sector included. For example, public transportation scales up to $2.5 + 2.5 \cdot 0.156 = 2.89\%$.

This results in an average carbon emission saved in the US of 100111704.6 metric tons C per year. This value is $\frac{105336765.5 - 100111704.6}{105336765.5} = 4.96\%$ less than the calculated value without perturbation. This difference can be explained by the fact that the most carbon emission saved comes from converting from cars to e-bikes, so reducing the number of cars converting to e-bikes will naturally decrease the amount of carbon saved.

For the United Kingdom, we do the same thing. The car percentage drops to $62.6 * 0.95 = 59.5\%$, a 3.1% drop. The rest of the sectors rise by $3.1/(100 - 62.6) = 0.0829\%$ every percent of sector included.

This results in an average carbon emission saved in the UK of 16,884,413.43 metric tons C per year. This value is $\frac{17751419.19 - 16884413.43}{17751419.19} = 4.88\%$ less than the calculated value without perturbation. This difference can once again be explained by the fact that the most carbon emission saved comes from converting from cars to e-bikes, so reducing the number of cars converting to e-bikes will naturally decrease the amount of carbon saved.

Traffic Congestion: To perturb the model, we increase and decrease the value of s_c by 5%. Reanalyzing the data, we get the following differences in values for commute times.

Table 4.4.1: Percent Change in Traffic Congestion Time for Changes in s_c

Country	$s_c + 5\%$	$s_c - 5\%$
US	6.14%	-6.01%
UK	6.27%	-6.13%

Health and Wellness: To perturb the model, we decrease the amount of car users by 5% arbitrarily, and increase the amount of other modes of transportation equally to make their total stay at 100%. We scale the P for each mode of transportation the precise same way as in the carbon emission perturbation.

Reanalyzing the data produces the following table:

Table 4.4.2: Perturbed Calories Burned by Transportation in US

Transportation	P	M	CTransport	WCalories
Car	71.8	35	0	452
Public Transportation	2.89	20	0	10.4
Walking	2.54	3.5	350	-7.29
Biking	0.462	15	750	-2.22
Taxicab/Other	1.73	20	0	0
Work From Home	20.7	0	0	0

Summing these values gives an AvgCalories = **453** calories. This value is $\frac{482-453}{482} = 6.01\%$ less from the calculated value without perturbation. This difference can be explained by the fact that the most calories gained from converting to e-bike is from cars. Thus, by reducing the number of cars switching to e-bikes, the average calories gained also is reduced.

For the United Kingdom, we do the same thing. The car percentage drops to $62.6 * 0.95 = 59.5\%$, a 3.1% drop. The rest of the sectors rise by $3.1/(100 - 62.6) = 0.0829\%$ every percent of sector included. Thus, the following table can be produced:

Table 4.4.3: Perturbed Calories Burned by Transportation in UK

Transportation	P	M	CTransport	WCalories
Car	59.5	35	0	375
Public Transportation	6.17	20	0	22.2
Walking	11.4	3.5	350	-32.7
Biking	3.57	15	750	-17.1
Work From Home	19.4	0	0	0

Summing these values gives an AvgCalories = **347** calories. This value is $\frac{369-347}{369} = 5.96\%$ less from the calculated value without perturbation. This difference can once again be explained by the fact that the most calories gained from converting to e-bike is from cars. Thus, by reducing the number of cars switching to e-bikes, the average calories gained also is reduced.

4.5 Strengths and Weaknesses

One strength of the model is that it is extremely flexible and scaleable with the population. Adjusting the percentages of the commuter populations for each mode of transportation is quite simple to do with our transparent formulas. The models for each of the factors are robust due to the consistent changes in the data from our sensitivity analysis.

A weakness of the model is that there is a lack of prior data on the percentage of people that switch from each transportation to e-bikes. Without understanding how people of various primary vehicular transportation react to the new technology of e-bikes, it is difficult to precisely represent the probability of a specific archetype transitioning to e-bikes.

5 Conclusion

5.1 Further Studies

Our first model fails to account for the adoption of e-bikes by those who are not regular bicycle users. By addressing this limitation, we can increase the accuracy of our predictions regarding the spread of e-bike technology throughout the United States and United Kingdom. Extending the Bass Diffusion Model to increase the market cap to accommodate the growth of bicycle users would also be fruitful. Furthermore, we can investigate further whether individuals who have purchased e-bikes are inclined to revert to conventional bicycles or other means of transportation.

Our second model had a very limited set of datapoints to work with, and given the black block nature of the random forest algorithm, we are unsure how the model was generated and how the importances were ranked. We can explore alternative modeling techniques that allow for more transparency and interpretability, helping us understand the models better.

Our third model was not able to fully encapsulate the effect of e-bikes on the three aforementioned factors due to the lack of data and many inferences that had to be made. In the future, we would like to further refine our model by gathering more comprehensive data and minimizing the need for assumptions. This would help us better understand and model how e-bikes impact traffic, individual health, and the environment.

5.2 Conclusion

In Part I, we predicted the number of e-bikes sold in two years and five years for both the United States and the United Kingdom. We created a Bass Diffusion Model and estimated the coefficients of innovation and imitation using a non-linear least squares regression. Subtracting the installed base fraction of two consecutive years and multiplying it by the market cap, we determined the predicted sales. In Part 2, urban population and electricity prices were estimated to be the most important in influencing people to choose e-bikes. Disposable income was a significant factor in the United States but not in the UK, while environmental perception was insignificant in both countries. In Part 3, we quantified the potential effects of transitioning to primarily e-bike transportation, focusing on the environment, traffic, and individual health. To estimate the likelihood of commuters adopting e-bikes, we factored convenience and cost, which enabled us to derive formulas for various functions and arrive at an approximate assessment of e-bikes' impact on the three factors.

6 References

1. Tacy, A. (2021, May 20). Bass diffusion model - Understand new product sales - SolvInnov. SolvInnov. https://solvinnov.com/innovation-diffusion-part-ii/

2. Survey: 100 million Americans bike each year, but few make it a habit. (2015, March 4). Streetsblog USA. https://usa.streetsblog.org/2015/03/04/survey-100-million-americans-bike-each-year-but-few-make-it-a-habit/

3. Cycling UK's cycling statistics. (n.d.). Cycling UK — The UK's cycling charity. https://www.cyclinguk.org/statistics

4. Ride Like the Wind, MathWorks Math Modeling Challenge 2023, https://m3challenge.siam.org/node/596.

5. "E-Bikes - U.S. Sales 2016." Statista, www.statista.com/statistics/326124/us-sales-of-electric-bicycles/.

6. Fact from fiction: Why consumers don't buy EVs : Blink charging. (2022, July 6). Blink Charging. https://blinkcharging.com/fact-from-fiction-the-real-reason-why-consumers-dont-buy-electric-vehicles/

7. Prices - Inflation (CPI) - OECD data. (n.d.). theOECD. https://data.oecd.org/price/inflation-cpi.htm

8. "Gridlocked Cities." Geotab, www.geotab.com/gridlocked-cities/.

9. "U.S. Census Bureau QuickFacts: United States." United States Census Bureau, 1 July 2022, www.census.gov/quickfacts/fact/table/US/PST045222.

10. "Can I Get a Tax Credit for My E-Bike Purchase?" HOVSCO, www.hovsco.com/blogs/blogs/can-i-get-a-tax-credit-for-my-e-bike-purchase. Accessed 6 Mar. 2023.

7 Appendix

7.1 part1.ipynb

```
# import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.optimize import curve_fit

$# Importing Data

# import TCP23_data.xlsx sheet Q1 E-bike Sales as df starting at row 6 and end
    at row 22
df = pd.read_excel('TCP23_data.xlsx', sheet_name='Q1 E-bike Sales', skiprows
    =6, nrows=17)

locations = ['US', 'Europe', 'France', 'China', 'India', 'Japan']

#rename rows to Year, US, Europe, France, China, India, Japan
df.columns = ['Year'] + locations

#focus on US and Europe data
df = df[['Year', 'US', 'Europe']]
locations = ['US', 'Europe']

# convert year data to int
df['Year'] = df['Year'].astype(int)

## Splitting Data by Location

# separate into data frames for each location with year
dataframes = {location: df[['Year', location]] for location in locations}

# for each df, drop all rows with no data
for location, df in dataframes.items():
    dataframes[location] = df[df[location] != '--']

# convert all values to numbers
for location, df in dataframes.items():
    dataframes[location][location] = pd.to_numeric(df[location])

# define market size for locations in 1000s of people
market_size = {'US': 45000, 'Europe': 40794, 'UK': 10700 }

# show us data head
print(dataframes['Europe'].head())

# divide data by market size
for location, df in dataframes.items():
    dataframes[location][location] = df[location] / market_size[location]

# for each dataframe create a column for cumulative sum
for location, df in dataframes.items():
    dataframes[location]['cum_sum'] = df[location].cumsum()

# for each dataframe create a new column with year minus first year
for location, df in dataframes.items():
```

```python
    dataframes[location]['year_diff'] = df['Year'] - df['Year'].iloc[0]

# show us data head
print(dataframes['Europe'].head())

## Defining Bass Diffusion Equation

# define bass diffusion model
def bass_diffusion_model(x, p, q):
    return (1.0- np.exp(-1.0 * (p + q) * x))/(1 + q / p * np.exp(-1.0 * (p + q) * x))

## Regression for US Data

location = 'US'

# get the last year in the location data
last_year = dataframes[location]['Year'].iloc[-1]

# get the first year in the location data
first_year = dataframes[location]['Year'].iloc[0]

# fit to data
bass_popt, bass_pcov = curve_fit(bass_diffusion_model, dataframes[location]['year_diff'], dataframes[location]['cum_sum'], p0=[0.003, 0.17], maxfev=100000)

# for sensitivity analysis
# increase p and q by 10% and -10%
bass_popt_10pplus = [bass_popt[0] * 1.1, bass_popt[1]]
bass_popt_10pminus = [bass_popt[0] * 0.9, bass_popt[1]]
bass_popt_10qplus = [bass_popt[0], bass_popt[1] * 1.1]
bass_popt_10qminus = [bass_popt[0], bass_popt[1] * 0.9]

# add columns for predictions
dataframes[location]['bass'] = bass_diffusion_model(dataframes[location]['year_diff'], *bass_popt)
dataframes[location]['bass_10pplus'] = bass_diffusion_model(dataframes[location]['year_diff'], *bass_popt_10pplus)
dataframes[location]['bass_10pminus'] = bass_diffusion_model(dataframes[location]['year_diff'], *bass_popt_10pminus)
dataframes[location]['bass_10qplus'] = bass_diffusion_model(dataframes[location]['year_diff'], *bass_popt_10qplus)
dataframes[location]['bass_10qminus'] = bass_diffusion_model(dataframes[location]['year_diff'], *bass_popt_10qminus)

# take results dataframe as Year, year_diff, and bass, bass_10pplus, bass_10pminus, bass_10qplus, bass_10qminus
US_results = dataframes[location][['Year', 'year_diff', 'bass', 'bass_10pplus', 'bass_10pminus', 'bass_10qplus', 'bass_10qminus']]

# create new df with years from last year+1 to 2030
preds = pd.DataFrame({'Year': range(last_year+1, 2030)})
preds['year_diff'] = preds['Year'] - first_year
preds['bass'] = bass_diffusion_model(preds['year_diff'], *bass_popt)
preds['bass_10pplus'] = bass_diffusion_model(preds['year_diff'], *bass_popt_10pplus)
preds['bass_10pminus'] = bass_diffusion_model(preds['year_diff'], *bass_popt_10pminus)
```

```python
preds['bass_10qplus'] = bass_diffusion_model(preds['year_diff'], *bass_popt_10qplus)
preds['bass_10qminus'] = bass_diffusion_model(preds['year_diff'], *bass_popt_10qminus)

# add preds to US_results
US_results = US_results.append(preds, ignore_index=True)

# plot results year against bass predictions and data
plt.plot(US_results['Year'], US_results['bass'])
plt.plot(dataframes[location]['Year'], dataframes[location]['cum_sum'], 'o', color='green')

#label the coordinate at x=2025 with (year, bass) to 6 decimal places
pred_2025 = US_results[US_results['Year'] == 2025]['bass'].iloc[0]
plt.plot(2025, pred_2025, 'o', color='red')
plt.annotate(f'(2025, {pred_2025:.6f})', xy=(2022, pred_2025))

#label the coordinate at x = 2028
pred_2028 = US_results[US_results['Year'] == 2028]['bass'].iloc[0]
plt.plot(2028, pred_2028, 'o', color='red')
plt.annotate(f'(2028, {pred_2028:.6f})', xy=(2025, pred_2028))

# set ticks to be every 2 years
plt.xticks(range(dataframes[location]['Year'].iloc[0], 2030, 2))

# label X axis as Year
plt.xlabel('Year')

# label Y axis as Proportion of Bike Market Using E-Bikes
plt.ylabel('Proportion of Bike Market Using E-Bikes')

# add legend with blue line for bass model, orange dots for original data, and red dots for predictions
plt.legend(['Bass Diffusion Model', 'Historical Data', 'Predictions'])

plt.show()
print("US Parameters: " + str(bass_popt))

# change in bass
US_results['bass_change'] = US_results['bass'].diff()
US_results['bass_10pplus'] = US_results['bass_10pplus'].diff()
US_results['bass_10pminus'] = US_results['bass_10pminus'].diff()
US_results['bass_10qplus'] = US_results['bass_10qplus'].diff()
US_results['bass_10qminus'] = US_results['bass_10qminus'].diff()

# find the percent change from bass to each of the sensitivity analysis basses
US_results['bass_10pplus'] = (US_results['bass_10pplus'] - US_results['bass_change']) / US_results['bass']
US_results['bass_10pminus'] = (US_results['bass_10pminus'] - US_results['bass_change']) / US_results['bass']
US_results['bass_10qplus'] = (US_results['bass_10qplus'] - US_results['bass_change']) / US_results['bass']
US_results['bass_10qminus'] = (US_results['bass_10qminus'] - US_results['bass_change']) / US_results['bass']

# add bikes sold as bass_change times market size
US_results['bikes_sold'] = US_results['bass_change'] * market_size[location]
US_results
```

```python

## Regression for Europe Data

location = 'Europe'

# get the last year in the location data
last_year = dataframes[location]['Year'].iloc[-1]

# get the first year in the location data
first_year = dataframes[location]['Year'].iloc[0]

# define bass diffusion model
def bass_diffusion_model(x, p, q):
    return (1.0- np.exp(-1.0 * (p + q) * x))/(1 + q / p * np.exp(-1.0 * (p + q
    ) * x))

# fit to data
bass_popt, bass_pcov = curve_fit(bass_diffusion_model, dataframes[location]['
    year_diff'], dataframes[location]['cum_sum'], p0=[0.003, 0.17], maxfev
    =100000)

# for sensitivity analysis
# increase p and q by 10% and -10%
bass_popt_10pplus = [bass_popt[0] * 1.1, bass_popt[1]]
bass_popt_10pminus = [bass_popt[0] * 0.9, bass_popt[1]]
bass_popt_10qplus = [bass_popt[0], bass_popt[1] * 1.1]
bass_popt_10qminus = [bass_popt[0], bass_popt[1] * 0.9]

# add columns for predictions
dataframes[location]['bass'] = bass_diffusion_model(dataframes[location]['
    year_diff'], *bass_popt)
dataframes[location]['bass_10pplus'] = bass_diffusion_model(dataframes[
    location]['year_diff'], *bass_popt_10pplus)
dataframes[location]['bass_10pminus'] = bass_diffusion_model(dataframes[
    location]['year_diff'], *bass_popt_10pminus)
dataframes[location]['bass_10qplus'] = bass_diffusion_model(dataframes[
    location]['year_diff'], *bass_popt_10qplus)
dataframes[location]['bass_10qminus'] = bass_diffusion_model(dataframes[
    location]['year_diff'], *bass_popt_10qminus)

# take results dataframe as Year, year_diff, bass, bass_10pplus, bass_10pminus
    , bass_10qplus, bass_10qminus
europe_results = dataframes[location][['Year', 'year_diff', 'bass', '
    bass_10pplus', 'bass_10pminus', 'bass_10qplus', 'bass_10qminus']]

# create new df with years from last year to 2030
preds = pd.DataFrame({'Year': range(last_year+1, 2030)})
preds['year_diff'] = preds['Year'] - first_year
preds['bass'] = bass_diffusion_model(preds['year_diff'], *bass_popt)
preds['bass_10pplus'] = bass_diffusion_model(preds['year_diff'], *
    bass_popt_10pplus)
preds['bass_10pminus'] = bass_diffusion_model(preds['year_diff'], *
    bass_popt_10pminus)
preds['bass_10qplus'] = bass_diffusion_model(preds['year_diff'], *
    bass_popt_10qplus)
preds['bass_10qminus'] = bass_diffusion_model(preds['year_diff'], *
    bass_popt_10qminus)

# add preds to results
```

```python
europe_results = europe_results.append(preds, ignore_index=True)
# plot results year against bass predictions and data
plt.plot(europe_results['Year'], europe_results['bass'])
plt.plot(dataframes[location]['Year'], dataframes[location]['cum_sum'], 'o',
    color='green')

#label the coordinate at x=2025 with (year, bass) to 6 decimal places
pred_2025 = europe_results[europe_results['Year'] == 2025]['bass'].iloc[0]
plt.plot(2025, pred_2025, 'o', color='red')
plt.annotate(f'(2025, {pred_2025:.6f})', xy=(2021, pred_2025))

#label the coordinate at x = 2028
pred_2028 = europe_results[europe_results['Year'] == 2028]['bass'].iloc[0]
plt.plot(2028, pred_2028, 'o', color='red')
plt.annotate(f'(2028, {pred_2028:.6f})', xy=(2024, pred_2028))

# set ticks to be every 2 years
plt.xticks(range(dataframes[location]['Year'].iloc[0], 2030, 2))

# label X axis as Year
plt.xlabel('Year')

# label Y axis as Proportion of Bike Market Using E-Bikes
plt.ylabel('Proportion of Bike Market Using E-Bikes')

# add legend with blue line for bass model, orange dots for original data, and
     red dots for predictions
plt.legend(['Bass Diffusion Model', 'Historical Data', 'Predictions'])

plt.show()
print("Europe Parameters: " + str(bass_popt))

# change in bass
europe_results['bass_change'] = europe_results['bass'].diff()
europe_results['bass_10pplus'] = europe_results['bass_10pplus'].diff()
europe_results['bass_10pminus'] = europe_results['bass_10pminus'].diff()
europe_results['bass_10qplus'] = europe_results['bass_10qplus'].diff()
europe_results['bass_10qminus'] = europe_results['bass_10qminus'].diff()

# find the percent change from bass to each of the sensitivity analysis basses
europe_results['bass_10pplus'] = (europe_results['bass_10pplus'] -
    europe_results['bass_change']) / europe_results['bass']
europe_results['bass_10pminus'] = (europe_results['bass_10pminus'] -
    europe_results['bass_change']) / europe_results['bass']
europe_results['bass_10qplus'] = (europe_results['bass_10qplus'] -
    europe_results['bass_change']) / europe_results['bass']
europe_results['bass_10qminus'] = (europe_results['bass_10qminus'] -
    europe_results['bass_change']) / europe_results['bass']

# add bikes sold as bass_change times market size
europe_results['bikes_sold'] = europe_results['bass_change'] * market_size[
    location]
europe_results
```

7.2 part2.py

```
# import libraries for random forest regressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error
from sklearn.model_selection import train_test_split
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

# Import data
df = pd.read_csv('input_data_us.csv')
# df = pd.read_csv('input_data_uk.csv')

# drop years before 2012 and after 2021
df = df[df['Year'] >= 2011]
df = df[df.Year <= 2021]

# Drop year
df = df.drop(['Year'], axis=1)

# create random forest model
rf_model = RandomForestRegressor(random_state=1)

# set target as 'sold' and features as everything else
y = df.sold
X = df.drop(['sold'], axis=1)

# fit model
rf_model.fit(X, y)

# get predicted values
rf_val_predictions = rf_model.predict(X)

# calculate mean absolute error
rf_val_mae = mean_absolute_error(rf_val_predictions, y)

print("Validation MAE for Random Forest Model: {}".format(rf_val_mae))

# plot predicted values vs actual values
rf_val_predictions

# plot predicted values vs actual values
plt.scatter(rf_val_predictions, y)
plt.xlabel('Predicted Values')
plt.ylabel('Actual Values')
plt.show()

# plot feature importance
importances = rf_model.feature_importances_
features = X.columns
plt.barh(features, importances)
plt.show()

# print feature importances in dict
feature_importances = dict(zip(features, importances))
print('feature importances:')
print(feature_importances)
```

```python
# Order the features by importance
feature_importances = dict(sorted(feature_importances.items(), key=lambda item: item[1], reverse=True))

# Assign each feature the number in the order
for i, key in enumerate(feature_importances):
    feature_importances[key] = i + 1

print('ranked features: ')
print(feature_importances)

### Perform sensitivity analysis

# Create empty dataframe with columns for each feature
sensitivity_df = pd.DataFrame(columns=X.columns)

print(X.columns)

## Sensitivity analysis dropping columns
for c in X.columns:

    # Drop column c
    X_copy = X.copy()
    X_copy = X_copy.drop([c], axis=1)

    # create new random forest model
    rf_model_new = RandomForestRegressor(random_state=1)

    # fit model
    rf_model_new.fit(X_copy, y)

    # Find new importances
    new_importances = rf_model_new.feature_importances_
    new_features = X_copy.columns
    new_importances = dict(zip(new_features, new_importances))

    # Order the features by importance
    new_importances = dict(sorted(new_importances.items(), key=lambda item: item[1], reverse=True))

    # Assign each feature the number in the order
    for i, key in enumerate(new_importances):
        new_importances[key] = i + 1

        # Find difference in importances
        new_importances[key] =  feature_importances[key] - new_importances[key]

    # Dropped column
    new_importances[c] = np.nan

    # Add new importances to sensitivity_df with pandas concat
    sensitivity_df = pd.concat([sensitivity_df, pd.DataFrame(new_importances, index=[c])])

print(sensitivity_df)


# Download csv of sensitivity_df
```

```
sensitivity_df.to_csv('sensitivity_df_us.csv')
# sensitivity_df.to_csv('sensitivity_df_uk.csv')
```