**Final Research Paper**

Chandler Via

Colorado State University Global

MIS581: Capstone – Business Intelligence and Data Analytics

Dr. Justin Bateh

October 2nd, 2022

**Table of Contents**

Abstract ………………………………………………………………………………………….. 3

Introduction ……………………………………………………………………………………… 4

Objectives ……………………………………………………………………………………….. 6

Overview of Study ………………………………………………………………………………. 7

Research Questions and Hypotheses ……………………………………………………………. 7

Literature Review ……………………………………………………………….……………….. 9

Research Design …………………………………………………………………………….….. 11

Methodology .…………………………………………………………………………... 11

Methods ………………………………………………………………………………… 12

Limitations ……………………………………………………………………………... 13

Ethical Considerations …………………………………………………………………. 13

Findings ………………………………………………………………………………………... 14

Conclusion ……………………………………………………………………………………... 19

Recommendations ……………………………………………………………………………… 20

References ……………………………………………………………………………………… 22

**First Draft of the Final Research Paper**

**Abstract**

This Capstone research project consists of an investigation into a rental bike company, Bay Wheels, which operates out of the San Francisco area. The main goal of the research is to build a foundation and guideline that can be followed to assess the growth and usage details of similar rental bike companies. Due to the highly limited scope of available data for these institutions, mainly detailing rentals within their system and locations, this research will stand as a guide to assist these companies in analyzing their own data. This research utilized a regression model to test for growth in Bay Wheels’ daily usage over time, spanning over the last 12 months. The regression model did prove that the platform was growing, successfully rejecting the null hypothesis that was presented. An Analysis of Variance test was also executed but was met with insufficient supporting evidence to successfully reject its appropriate null hypothesis. This test was utilized in the hope to prove that particular groups around the city received more use than others. This area stands as an area of possible review or retesting with a different strategy.

**Introduction**

Combustion-powered vehicles mark a core attribute of American Society for the past 8 decades. The presence of these vehicles, as Metz (2013) explains, is due to the U.S. involvement in WWI and the effects of the Great Depression on the automobile industry. Electric vehicles and horses, the two alternatives to gasoline powered automobiles, proved too difficult and time consuming for continued use in the field over extended periods and speeds. After the war, these automobiles were inserted into American society on a grand scale, which influenced the infrastructure to mirror their needs quickly. Soon after, the Great Depression caused a myriad of different manufacturers to forgo their electric vehicle lines in favor of the more profitable gasoline models instead. These dual reasons solidified the usage of gasoline powered automobiles in U.S. life.

Pollutants are one of the major concerns about the continued use of gasoline powered automobiles. Of these pollutants, one of the most heavily discussed is CO2 due to the evidence that the growing presence of it has contributed to hotter environmental temperatures, which contributes to more energy use to combat the uncomfortable heat (Solomon, Plattner, Knutti, & Freidlingstein, 2009). The United States Environmental Protection Agency (2022) has noted that the US creates 8,887 grams of Carbon Dioxide for every gallon used by the average automobile. This allows for a quantifiable amount of pollution that is created per trip taken by an individual. On a mass scale, the overwhelming presence of automobiles has accelerated the appearance of CO2 in the atmosphere. With this information, many recommend that a reduction in vehicle use is preferable, as it reduces the total emissions created, and can help return to stable planet temperatures.

The primary recommendation for alternative transportation is the usage of bicycles instead. While a simple choice, many people find issue with the logistics of owning, securing, and maintaining a bicycle. This concern has generated a new system that has grown over the past half-decade: E-bikes. E-bikes are rentable bicycles, both electric and standard, that can be quickly retrieved from a docking station, rode to another location, and then stored in a similar docking station. These E-bikes are maintained by a multitude of different companies across the nation and take up the burden of logistics for stocking and maintaining these bicycles across their respective city. Switching to this alternative method not only removes the potential increase in pollution from the same trip taken by an automobile, but it also improves the user’s health with the physical activity. These rental companies also offer electric models that use battery power to assist in the continued movement after the user initially starts from a resting position. These options allow for users to quickly move to their desired location, and are benefitted by cities that already have allotted bike lanes on the road.

One issue that these companies must be concerned about is the rate of change of customers who utilize their platform. While growth is seen as a positive aspect, they must ensure that these new users do not interfere or diminish the customer experience of more veteran users. This growth is not only beneficial for the revenue for the company operating the system, but also for reducing the number of trips that are taken with gasoline powered vehicles. This interaction requires the company to maintain and focus on the customer experience and availability of not only the E-bikes, but also the available open docking stations in the user’s chosen destination. This growth rate must be examined, and areas of interest must be found to allow the company reasonable expectations for their expansion operations. For this research, the Bay Wheels company, an organization operating in the San Francisco area, will be analyzed.

**Objectives**

The primary focus of this research is to identify and prove a rate of growth has been achieved for the platform across the past 12 months. By viewing the number of daily rentals across the entire population, a growth should be able to be identified and reported back to the company that the user base has expanded in use over time. The time will span from August of 2021 to July 2022. Daily usage counts may be too specific to achieve a competent test result, so alternative granularities will be applied if this case appears. Any results that are less than or equal to zero would indicate either stagnation or a recession of active users to the platform.

While growth is important to the platform, the identification of high-frequency areas of use is also important. The main purpose of the second element to this research problem seeks to identify these key areas. Areas of high use should be identified to allow the company to focus on priority locations, to better accommodate for the expansion of the platform. By targeting frequently used stations, Bay Wheels can adjust or add additional stations for more E-bikes to be stored or retrieved in those locations. This can help prevent interference between new users to the platform and those that are more established, all the while providing a more confident platform for all.

By investigating these two elements, the company can have a better insight into their own system to better accommodate to this growth, and can even use these statistics in conjunction with environmental facts about vehicle pollution to push for stronger advertising. These two objectives can also be utilized as a guideline for similar research to be accomplished at similar organizations as well. These tests and steps could allow the company to persuade more users to join their platform, given the societal benefit their alternative transportation provides for the community.

**Overview of the Study**

Lyft (2022) publishes the Bay Wheels transactional trip data on a monthly basis. Each trip taken by an individual is logged into this data, anonymized, and then released to the public monthly. These details include the starting and ending station location, ID, and coordinate data, along with information about trip time and bike type. By exploring the data present in this repository, the objectives previously highlighted would potentially be possible to complete.

This Capstone project will start with the defining of research questions to promote insights about the discussed objectives, followed by further establishing of hypotheses that correspond to those questions. These hypotheses will consist of a null and alternative hypothesis for each question, along with guidelines for rejection of those null hypotheses. A literature review will be discussed next, adding context to the subject matter, along with the research design and architecture. Finally, the project will conclude with a discussion of the individual results of the research, followed by a conclusion and possible recommendations that could be pursued from this research.

**Research Questions and Hypotheses**

The two goals of this research will be utilized to identify specific research questions that can be used to aid direction. The primary goal is to establish a growth of the platform over time, in terms of daily rentals as the year progresses. To write this subject as a question, “Did the company experience a growth in daily rentals across the period of time?” This question derives from the desire to find how well the company has been increasing in terms of use on the platform. Active user counts, although tangentially related, will not give a certain understanding of how often the platform is utilized over time. The secondary objective, determining stations with high frequency, will be reiterated into the following question as well: “Are there stations that receive more daily use than other stations?”. By answering the first question, the second question can further illuminate the company on next steps towards accommodating new users.

With the two business questions identified, it would be pertinent to further define hypotheses for each one that can be tested to prove their factualness. For the first question, regarding platform growth, the following null hypothesis will be used, “H0: There is no evidence of a relationship between cumulative daily rentals and time.” To go along with this, the following alternative hypothesis will be used, “Ha: There is evidence of a relationship between cumulative daily rentals and time.” This hypothesis can be tested, and if the null hypothesis is found to be rejected, a manual view of the slope or coefficient will establish a positive or negative rate of change.

Identification of key stations that receive more use than others remains a core component of the second business question. The null hypothesis that matches this would be, “H0: There is no evidence of a difference in average daily use between stations.” The alternative hypothesis to match this would be, “Ha: There is evidence of a difference in average daily use between stations.” Due to complications with population variance and category size, these stations may have to be clustered together to promote a better understanding of user tendencies. Kim (2017) concretes this point by highlighting many of the issues that can appear when the number of categories expands within a test. Of those issues, one of the most heavily scrutinized is the over abundance of categories available in an ANOVA test. This problem can skew the results and can over-favor elements that simply have a heavier presence. This point will be discussed further in detail in the methodology section.

**Literature Review**

Regarding the subject of literature reviews, Adams et al (2014) states that issues in similar research should be discussed to better illuminate the audience on the context behind the subject. Previous conclusions and inferences are also beneficial because they provide a wider understanding of the subject matter. The literature review will begin with a discussion on gasoline vehicle presence in US markets, followed by their effects on climate and potential alternatives. Lastly, research surrounding user tendencies with bicycles will be highlighted to illuminate how these growing cities could make changes to promote bike usage.

**Example 1 – History of Gasoline Powered Vehicle Usage**

Metz (2012) offers a myriad of reasons as to the main forces that influenced how ingrained gasoline powered vehicles are in American life. The two main points, to recover the points from the introduction of this research, are the influence of WWI and the Great Depression around the turning of the century. These two impactful events dictated a degree of necessity on these vehicles and facilitated the removal of many other alternatives. This point is reinforced with details given by Moody et al. (2021), where they estimate that roughly 91% of adults in the U.S. require and use a car to reach their workplace.

In the past decade, the reappearance and popularity of alternative transportation has risen. Popular movements in favor of public transportation, electric vehicles and the resurgence of bicycle use are commonly found online. With the recent improvements made in battery and power storage, electric vehicles and new, electric bikes are viable modes of transportation that can compete with the gasoline vehicle industry. This change, along with the appearance of more local options for personal necessities, has streamlined a period of American culture that can now facilitate an active choice between car and alternative transportation.

**Example 2 - Gasoline Powered Automobile Pollution**

Gasoline powered vehicles have been proven to generate pollutants that negatively affect the nearby ecosystem for both planet and human health. Johansson et al (2017) utilized previous research that quantified human life loss from local air pollutants, and estimated a total loss of life that could be prevented by switching a number of commuters from car to bicycle. They specified users that live within a 30 minute ride from their home to place of work, which totaled 111,487 people that fit this criteria. If they were to change their transportation methods to bicycle, they would prevent a total loss of 450 years of life from the reduction in pollutants caused locally. This came from an estimate from Johansson et al, where they found that 6% of nonaccidental mortality is found per 1 μgm-3 of carbon nearby.

Similarly, Southerland et al. (2021) established similar research in the same locale as the Bay Wheels system, San Francisco’s Bay area. They found that over 2,500 deaths could be attributed to Nitrous Dioxide pollutants, along with 150 deaths related to Black Carbon. Extrapolating these details, similar to the research done by Johansson et al (2017), one could estimate the amount of life loss prevented by a large shift in transportation method. This loss of life can be multiplied by the standard presence of people on the streets and in the central city area. These details make a compelling argument for Bay Wheels and the community to request the city add more bike friendly infrastructure and safety, which will encourage more users to make the switch away from gasoline vehicle use.

**Example 3 – Bicycle Infrastructure**

Other groups have completed research similar to this Capstone project. Singhvi et al (2015) established research in this way, attempting to predict the usage of another E-bike platform using historical data on the platform’s use, weather data and taxi information. They managed to create a model that can predict the usage of the platform a couple days ahead of the current date. This stands as an example of some of the research done on these systems and how they tend towards being specific for an individual company and city. Singhvi et al aimed to utilize prediction strategies for nearby dates with high accuracy, while this research project aims to find a rate of change of use in the platform across an entire year. This remains an excellent example, mainly from it’s display of strategy and analytics that could be similarly preformed on another E-bike company.

**Research Design**

**Methodology**

The Bay Wheels data set contains a transactional oriented organization, with each row corresponding to an individual rental instance. Each rental logs information about the start time, end time, the bike type used in the rental, and much more. The data will be cleaned before analysis to ensure that the test are not affected by missing values. These missing values will be briefly examined as well to ensure that they are missing at random, and do not constitute categorically missing values. If not, they will be entered as their own category. This process will begin with a brief exploratory search through the data and then will follow with the hypothesis testing.

To test the hypotheses, two statistical tests will be performed to assist in the rejecting of the null hypotheses of both. A regression function will be utilized to test for the first hypothesis, specifically in an attempt to identify a rate of growth over time. Following this, an Analysis of Variance (ANOVA) test will then be used to test for a higher frequency of rentals from some stations than others. Given that a regression model will result in a coefficient of change and an ANOVA test will identify the magnitude to which a group is used more than others, the research would primarily be considered quantitatively oriented.

**Methods**

To correctly test for the first hypothesis, “H0: There is no evidence of a relationship between daily platform use and date”, a regression model will be created. Linear regression maintains a handful of assumptions that must be confirmed before the results of the model can be considered. Schmidt & Finan (2018) give more context to the topic, and state that the assumption for normality should not be stressed as heavily if it is not met. It is much more damaging to apply transformations to make the data fit, so this will be avoided for the matter of the regression test. Using the details provided by Schneider, Hommel & Blettner (2010) on regression hypothesis rejection, the null hypothesis can be rejected if the coefficient for the predictor variable is not equal to zero. For the matter of the business question associated with this hypothesis, a positive value for the coefficient is preferred.

Kozak & Peipho (2017) detail the assumptions that should be met when using an ANOVA test and will be used as a guideline for this element of the research. The assumptions are as follows: The population is distributed normally, independence of the variables, and homogeneity of variance. If these assumptions are not met, alternative strategies will be utilized to attempt to be sufficient for the test.

**Limitations**

The biggest consideration for this data’s limitations would be the defined scope of location and company. Any insights gleaned from this research will be unable to be applied to any company or location outside of the Bay area and the Bay Wheels company. Even if the researcher were able to conglomerate a master list of all rentals taken by users across the nation, across companies and systems, these tests would not be able to easily be applied. The ANOVA test, for instance, would suffer massively from the magnitude of different categories present. However, this research can stand as a framework for further research into other organizations and systems around the country. By following these steps, they can recreate and analyze their results individually to assess trajectory.

Another limitation that is present is the extremely anonymized data available. Because there is no information about the customer themselves, further insights about pattern and tendencies are unable to be investigated at this time. An example of this would be residential area or address so that a distance from starting location could be assessed. Similarly, analyzing common ending destinations (Like work places, shopping centers, etc.) would give insights on where new station locations would be beneficial to those users. If these more intricate pieces were available, the number of potential insights would be exponentiated.

**Ethical Considerations**

The anonymized nature of this data is a boon regarding ethical considerations in data. However, due to the transparency of Bay Wheel’s rental information and locations, rival companies can utilize this data to capitalize on any aspects that might be lacking or in development for Bay Wheels. This increases the power for competitors past those that normally would not release this information to the public. Similarly, investors to the company that are vigilant might notice a downtrend dramatically in the rental usage, causing negative impressions on data that might be relatively incomplete due to the anonymized visibility. Investor confidence, or lack thereof, can affect the platform itself and could cascade to reduced use and, in turn, more users returning to pollutant-heavy transportation options.

**Findings**

The assumptions of linear regression must be assessed before the results of the model can be accepted and internalized. SAS Studio was leveraged to create the regression model, which generates residuals statistics and graphs that can be used to analyze the data’s consistency and fit. Figure 1 shows one of the most important of these graphs, the Q-Q plot. The Q-Q plot shown in the first column, second row, shows a fairly strong fit of the data against the central axis. Casson & Farmer (2014) explain that this denotes an acceptable normality in the variables. Furthermore, they explain that the homoscedasticity is equally important, and is weaker than the Q-Q fit shown previously. Figure 1’s second chart shows points that reside outside of the acceptable range of -2, 2. This range is specifically for the RStudent value and lacks symmetry, especially given the large decrease found in the month of December to January. This slightly violates the assumption, but not enough to completely invalidate the model.

**Figure 1**

Fit Diagnostic Graphs for the Regression Model Between Daily Use Versus Date

Diagram, engineering drawing

Description automatically generated

The regression model’s statistics resulted in a significant measure for the P value. This result can be viewed in Figure 2, located within the second table. The model’s significance is diminished by the low R-squared value found in the third table, with a value of 0.0892. This translates to only a 8% influence in variance for the predictor variable (date) against the response variable (daily rental count). The second area to focus on within Figure 2 is the coefficient present for the “started\_at\_date” variable. This coefficient of ~4.7 satisfies the part of the business question associated with this test to investigate a potential positive growth of use of the platform. This significant test result gives enough support to reject the null hypothesis of “H0: There is no evidence of a relationship between platform daily use and data” and the positive rate of change gives enough support to conclude that the platform is increasing over time.

**Figure 2**

Regression Model Statistics for the Daily Use Against the Started At Date Variable



While the fit graphs and the statistics table are sufficient to analyze the regression model itself from a technical standpoint, it is still a wise choice to view the fit plot for the regression model. Figure 3 shows this fit plot and details the daily count of individual rentals against the date. The trend line shows a slightly upwards slope that grows throughout the year. This upwards momentum mirrors the positive coefficient that was discussed previously. It should be noted that there is a large decreased period of use between December to February, which is presumed to be due to the cold weather that is present in San Francisco in that period.

**Figure 3**

Fit Plot for the Daily Use versus Date Regression Model

Graphical user interface, chart, scatter chart

Description automatically generated

The second hypothesis involved an ANOVA test against different groups of stations across the Bay area. As stated in previous sections, an ANOVA test would be vulnerable against a multitude of categories, so reducing the vast number of stations into groups would be the wisest course of action. To do this, the data was imported into R and clustered using the K-means algorithm. 5 groups were determined, two for the San Francisco downtown area, two for the Nappa Valley area, and one for the San Jose area. These groups can be seen below in Figure 4, a visualization of the groups based on their coordinate data, generated with R-studio.

**Figure 4**

R Code Used to Cluster and Visualize the Groups of Bike Stations for the Bay Wheels Data

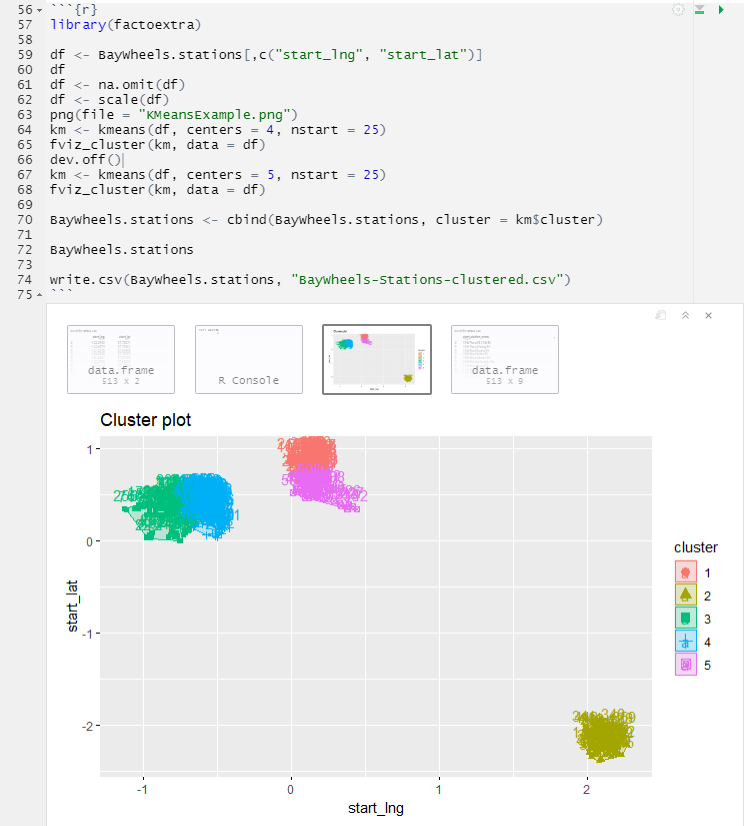
****

Figure 5, below, shows the resulting tables from the ANOVA test. These tables highlight a statistically significant model, regarding the P value, and a mildly strong R-squared value of ~0.384. Although this information is held in positive regards to the matter of rejecting the null hypothesis, the Levene’s test shown in the bottom of Figure 5 would refute this point. Levene’s test, when significant, indicates a violation of homoscedasticity among groups. Due to this, the positive results of the ANOVA cannot be utilized.

**Figure 5**

ANOVA Test Statistic Tables for Cluster Group versus Average Daily Use

Graphical user interface, application

Description automatically generated

Efforts were made to circumvent this null result, but none were successful. An investigation was completed in the direction of a Welch’s ANOVA test, but this matter would only suffice if the data was found to be normal with a significant Levene’s test. In this case, the data was not found to be normal, and thus rendered the Welch’s ANOVA test inept. This matter will require further research and development to glean useful insights.

**Conclusion**

The first hypothesis, “H0: There is no evidence of a relationship between platform daily use and date”, was rejected due to the significant Regression model and compounding assumptions. The alternative hypothesis, “There is evidence of a relationship between platform daily use and date”, can be accepted and the positive coefficient found indicates that the platform is, indeed, growing. This result indicates to the company that the platform is moving in a positive direction, and concerns about platform growth are worthy of consideration, albeit at a slow rate.

The results for the ANOVA test were hopeful but lacking in needed support to allow for statistical use. In this vein, the null hypothesis of “H0: There is no evidence of a difference in average daily use between station groups” cannot be rejected. This means that the business question of “which areas/stations receive higher use than others” could not be answered at this time. This area was explored for potential, quick alternatives to solve the hypothesis at hand, but none were found that could satisfy the requirements of the question along with the statistical support needed for confidence.

In summary, the testing was able to conclude half of the business questions that were set forth. Of those business questions, the first was used as the foundation for the second and was proven to be successful, which should allow for subsequent testing to follow it’s work without issue. Alterations to the ANOVA test could be potentially performed, with transformations applied to the data at hand or by filtering the data to exclude areas like San Jose. Similarly, other tests could also be implemented to answer more business questions, branching out from the original regression results.

**Recommendations**

This project has a multitude of areas that could be expanded on to develop more insights for the Bay Wheels company. The first of which involves the exploration of time data that is available in the data online. While running the regression model, a temporary downward trend was found during the December to February period. This winter period would correspond with more harsh weather, and a downtrend in use of the Bay Wheels system. This avenue could be investigated further to develop seasons where maintenance can be prioritized due to the disuse of the bicycles in the period. Similarly, by minimizing the amount of bikes lingering in stations around the area, weather damage can be mitigated during that time as well through careful analysis.

Another area that could use potential investigation would be the high-frequency areas that are used the most. A total count for the year was easily generated during these tests and this count could be easily sorted for the most used locations in the system. However, the objective stated was to investigate the groups of stations that receive a higher number of uses on a daily basis. These locations would be the ones to benefit first from upgrades or expansions to the system**.** It is possible that the locations present in the San Jose area adversely affected the group analysis, primarily due to the low use which could influence the corresponding Levene’s test. This avenue would be an excellent opportunity for future researchers to follow up and unlock an insightful area of this data.

**References**

Adams, J., Khan, H. T. A., & Raeside, R. (2014). *Research Methods for Business and Social*

*Science Students.* Sage Publications.

Casson, R. J. & Farmer, L. D. M. (2014). Understanding and Checking the Assumptions of

Linear Regression: A Primer for Medical Researchers. *Clinical & Experimental*

*Ophthalmology. 42(6), 590-596*. <https://doi.org/10.1111/ceo.12358>

Johansson, C., Lovenhiem, B., Schantz, P., Wahlgren, L., Almstrom, P., Markstedt, A.,

Stromgren, M., Forsberg, B., & Sommar, J. N. (2017). Impacts on Air Pollution and

Health by Changing Commuting from Car to Bicycle. *Science of the Total Environment.*

*584, 55-63.* <https://doi.org/10.1016/j.scitotenv.2017.01.145>

Kim, T. K. (2017). Understanding One-way ANOVA Using Conceptual Figures. *Korean Journal*

*of Anesthesiology. 70(1), 22-26.* <https://doi.org/10.4097%2Fkjae.2017.70.1.22>

Kozak, M. & Peipho, H. P. (2017). What’s Normal anyway? Residual Plots are More Telling

than Significance Tests when Checking ANOVA Assumptions. *Wiley Journal of*

*Agronomy and Crop Science. 204, 86-98.* https://doi.org/10.1111/jac.12220

Lyft. [n.d.]. *System Data*. <https://www.lyft.com/bikes/bay-wheels/system-data>

Metz, D. (2013). Peak Car and Beyond: The Fourth Era of Travel. *Transport Reviews. 33(3),*

*255-270.* <https://doi.org/10.1080/01441647.2013.800615>

Moody, J., Farr, E., Papagelis, M., & Keith, D. R. (2021). The Value of Car Ownership and Use

in the United States. *Nature Sustainability.* https://doi.org/10.1038/s41893-021-00731-5

Singhvi, D., Singhvi, S., Frazier, P. I., Henderson, S. G., Mahoney, E. O., Shmoys, D. B., &

Woodard, D. B. (2015). Predicting Bike Usage for New York City’s Bike Sharing

System. *Association for the Advancement of Artificial Intelligence.*

<https://people.orie.cornell.edu/pfrazier/pub/2015_SinghviSinghviFrazierHenderson.pdf>

Schmidt, A. F. & Finan, C. (2018). Linear Regression and the Normality Assumption. *Journal of*

*Clinical Epidemiology. 98, 146-151.* [https://doi.org/10.1016/j.jclinepi.2017.12.006](https://doi-org.csuglobal.idm.oclc.org/10.1016/j.jclinepi.2017.12.006)

Schneider, A., Hommel, G., & Blettner, G. (2010). Linear Regression Analysis. *Deutsches*

*Arzteblatt International . 107(44), 776-782.* <https://doi.org/10.3238%2Farztebl.2010.0776>

Solomon, S., Plattner, G. K., Knutti, R., & Friedlingstein, P. (2009). Irreversible Climate Change

Due to Carbon Dioxide Emissions. *Proceedings of the National Academy of Sciences.* <https://doi.org/10.1073/pnas.0812721106>

Southerland, V. A., Anenberg, S. C., Harris, M., Apte, J., Hysad, P., van Donkelaar, A., Martin,

R. V., Beyers, M., & Roy, A. (2021). Assessing the Distribution of Air Pollution Health

Risks within Cities: A Neighborhood-scale Analysis Leveraging High-Resolution Data

Sets in the Bay Area, California. *Environmental Health Perspectives. 129(3).*

<https://doi.org/10.1289/EHP7679>

U.S. Environmental Protection Administration. (2022). *Greenhouse Gas Emissions from a*

*Typical Passenger Vehicle.* https://www.epa.gov/greenvehicles/greenhouse-gas-

emissions-typical-passenger-vehicle