**Business Intelligence and Data Analytics Capstone – Option 1**

Ryan Meyers

Colorado State University Global

MIS581: Capstone – Business Intelligence and Data Analytics

Dr. Justin Bateh

December 5, 2021

**Contents**

[Abstract 3](#_Toc89525624)

[U.S. Organization Background 4](#_Toc89525625)

[Ford Motor Company History 5](#_Toc89525626)

[Products and Services 6](#_Toc89525627)

[Locations and Employee Size 6](#_Toc89525628)

[Revenue 7](#_Toc89525629)

[Dataset Description 7](#_Toc89525630)

[Types of Variables 8](#_Toc89525631)

[Visual Model of Data – Data Dictionary 9](#_Toc89525632)

[Utilization of the Dataset 10](#_Toc89525633)

[Objectives 11](#_Toc89525634)

[Overview of the Study 12](#_Toc89525635)

[Overview of Hypotheses 12](#_Toc89525636)

[Research Hypothesis 14](#_Toc89525637)

[Literature Review 15](#_Toc89525638)

[Common Challenges in Security, Privacy, and Ethics of Data 17](#_Toc89525639)

[Security 18](#_Toc89525640)

[Privacy 18](#_Toc89525641)

[Ethics 19](#_Toc89525642)

[Research Design 20](#_Toc89525643)

[Methodology 20](#_Toc89525644)

[Methods, Tool, and Techniques 21](#_Toc89525645)

[Limitations 22](#_Toc89525646)

[Ethical Considerations 22](#_Toc89525647)

[Data Analysis and Findings 23](#_Toc89525648)

[Conclusion 38](#_Toc89525649)

[Recommendations 39](#_Toc89525650)

[References 41](#_Toc89525651)

# Abstract

The research conducted in this report utilized data exploration and analysis techniques to develop a framework for statistical inferences on a public dataset in the automobile industry. The researcher developed descriptive and predictive statistical models to investigate the relationship of variables collected in the Environmental Protection Agency’s (EPA) Automobile Trend dataset. These relationships were tested for statistical significance in terms of predictive capabilities on generating accurate value predictions for future automobile miles-per-gallon figures. The methods for producing these predictive models and common challenges in data analytics were explored in this report, in addition to, explanations of popular models in predictive data analytics (linear and time-series regression). The research found that statistically significant models could be developed using the EPA’s automobile dataset and the researcher provided the impact that these results could have for a United States based company: Ford Motor Company. The report concludes with recommendations for further analysis and the outlook on the automobile industry into the future.

*Keywords*: Statistics, predictive models, linear regression, Ford Motor Company, EPA

**Business Intelligence and Data Analytics Capstone – Option 1**

The proliferation of data in recent years through the datafication of various forms of individual life, as well as, the advancements in storage and collection capabilities has created a data rich environment for organizations. The ability to harness information from these various data sources yields the possibility to gain competitive advantages that can be leveraged from the data know available to any specific company. This report will highlight the research methods and design that will include the tools and techniques used to analyze a public dataset. The public dataset will assist the selected United States organization in demonstrated how data analytics can be leveraged for industry specific insights into the current performance and future predictions of the product markets. The selected organization is Ford Motor Company so the analysis will focus on the automobile industry utilizing a public dataset provided by the United States Environmental Protection Agency for automotive trend data.

# U.S. Organization Background

The company selected for this report and the capstone analysis is Ford Motor Company. Ford operates within the automotive industry offering a variety of products and services to customers internationally while being headquartered in the United States. This section of the report will cover the history, products/services, facility locations, employee size, and revenue/financials of the organization. The researcher selected Ford because of the company’s manufacturing background and expertise. Ford will provide a solid basis for conducting analytical research on a company in the manufacturing and supply chain field, that the researcher currently works within, so that the tools applied to the capstone project can also be applied in the researcher’s area of professional work and interest (Wee & Wu, 2009). Ford Motor Company is also a leader in the innovative space of the industry so new information gathered from research will assist the organization in a more impactful way than other companies in the automobile industry.

## Ford Motor Company History

Ford has been recognized as one of the automotive giants in the global market place from its early years as a company. The Ford Motor Company was formally founded in the year 1903 by Henry Ford and a few key stakeholders in the city of Detroit, Michigan (“Ford Motor Company Incorporated*,* 2009). Henry Ford had a few failed attempts prior in developing and solidifying a car manufacturer prior to the creation of the motor company that takes his name; Ford. The first assembly plant for automobiles was established in Detroit on Mack Avenue where the company was originally founded. At this time, the company was building cars manually and eventually the company grew to be the largest U.S. car manufacturer in 1906 producing 8,729 cars in that year (“Timeline - History, Chronology - Ford Rouge Factory Tour, 2021). The company continually grew when the popular Model T, which was an affordable and reliable car, was produced in the year 1908. Ford Motor Company saw the development and creation of multiple different American car companies over the years since the creation of the Model T to the year 1925 when Chrysler Motor Corporation was founded creating three major car companies in the United States: Ford, General Motors, and Chrysler. Ford was a continual innovator developing the assembly line production system to mass produce automobiles and is one of the largest companies in the world in terms of assets and wealth. The company is currently headquartered in Dearborn, Michigan.

## Products and Services

Ford Motor Company has a long history of owning various car manufacturing brands under the Ford corporate structure. The organization has owned car producers such as Jaguar, Volvo, and Mercury, however, the company currently only has three offerings, Ford, Lincoln, and Troller. This was the result of the company selling a large portion of the brands over time to meet market needs or focus on internal competencies. Ford and Lincoln offer a variety of products under these brand names including sedans, trucks, luxury cars, sport utility vehicles, etc. The product offerings have changed with market demand such as larger demand for trucks in the United States compared to cars by U.S. companies. Ford also offers financial services for those purchases through their locations called Ford Credit Company.

## Locations and Employee Size

Ford Motor Company has expanded through various different countries from their beginnings in Detroit, Michigan. The company is well-known for the plant built in Dearborn, Michigan (which is still in operation) shortly after the creation of the organization but the manufacturing facilities have developed since those initial plants. Ford has factories in over 125 locations globally with some being entirely owned by Ford Motor Company and others being join-venture facilities (“Locations”, 2021). This is not inclusive of the multitude of Ford and Lincoln dealers that are found internationally selling Ford automobiles. These various facilities employ around 186,000 employees globally with 57,000 of those employees being hourly workers in the U.S. and 88,000 total U.S employees.

## Revenue

Ford’s annual revenue as of August 2021 amounted to $127.1 billion with profits taking a $-1,279 million decline (“Ford Motor: 2021 Fortune 500”, 2021). This allows Ford Motor Company to be the largest U.S. based automobile manufacturer, rising past General Motors, and the fourth largest automobile manufacturer in the world. The stock price for Ford Motor Company is currently at $15.70 a share as traded on the New York Stock Exchange (NYSE).

# Dataset Description

The automobile industry has a variety of datasets that could be useful to organizations seeking a competitive advantage or new knowledge in the ever-evolving market climate. The dataset selected for the capstone analysis is provided by the United States Environment Protective Agency (EPA). The dataset is provided from the Trends database that has been the work of the EPA since 1975 to the current year. The dataset will contain data concluding in the year 2020 since the year 2021 is on-going so the scope of the analysis will contain information from 1975 to 2020. The data is delivered in a comma-separated values (csv) file exported from the EPA database. The variables contained within the dataset include: model year, vehicle type, government regulatory class, production share, miles per gallon, weight, horsepower, CO2 emissions, etc. The EPA Automotive Trends dataset will provide the basis for the data used in the capstone analysis. The researcher will also leverage other datasets on automotive sales from the organization called Knoema which is a source company for data sources. The datasets that will be used from this organization include: *Top Vehicle Manufacturers in the U.S. Market* and *The Global Electric Car Market* datasets. The data includes is sales figures for the top vehicle manufacturers each year and the expansion of new technology automobile sales such as electric cars which is data found within the Department of Energy’s published data. The links to the dataset sources can be found in *Table 1*.

**Table 1**

Links to dataset sources.

|  |  |
| --- | --- |
| **Name** | **Website Link** |
| EPA Dataset | https://www.epa.gov/automotive-trends/explore-automotive-trends-data |
| Top Manufacturers | https://knoema.com/infographics/floslle/top-vehicle-manufacturers-in-the-us-market-1961-2016 |
| Electric Vehicles | https://knoema.com/infographics/ffgfqj/the-global-electric-car-market |

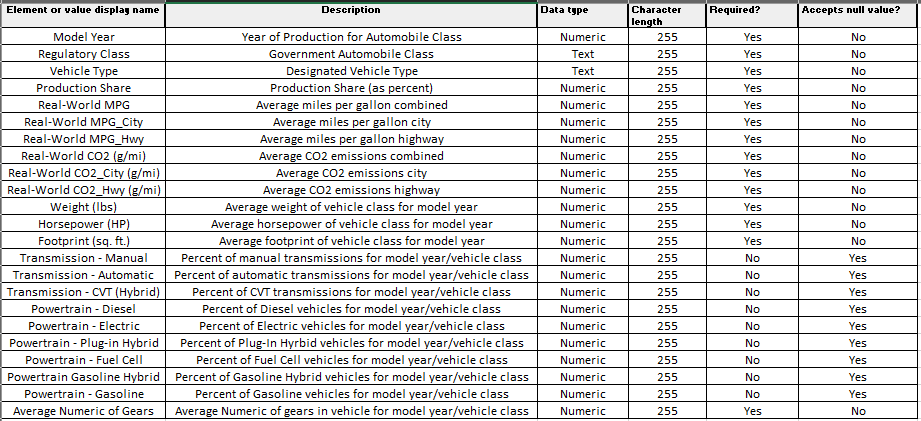
# Types of Variables

Datasets can contain a multitude of different variable types for the different variables. The main types are described as categorical, ordinal, interval, and binary variables. Categorical variables (referred to as nominal variables) are variables that fall within a specific category but the ordering of the categories is not significant. A binary variable is included within the categorical classification, in which, there are two categories of yes or no (1 or 0). Categorical variables can also have more than two categories for a specific variable. Ordinal variables are similar to categorical variables in that there are more than two categories for the variable but order is significant. For example, rankings in pain where the difference in rankings is meaningful but the difference in scale is not exact. Interval variables (or numeric variables) are variables where a scale is applied and the difference between values in the scale is known for the variable. Ratio variables are another type of variable which is the same as interval variables, however, the value of zero is known and relative to the other variables (Kaur et al., 2018). The EPA dataset in this report uses a mixture of the variables such as regulatory class of automobiles falling within the categorical variable type and real-world mpg under the ratio variable type.

# Visual Model of Data – Data Dictionary

The creation of a data dictionary allows for documentation that can be used to understand the variables in a dataset. “A data dictionary is used to catalog and communicate the structure and content of data, and provides meaningful descriptions for individually named data objects” as defined by the U.S. Geological Survey organization from the U.S. Department of the Interior (2021). The data dictionary provides context to the main public dataset that is being used by the researcher in a visual tabular format for quick reference by the researcher and the audience of this report. The researcher has provided a listing of the variables in the proposed main dataset and attributes of these variables as shown in *Table 1.*

**Table 1**

Data Dictionary for the EPA dataset of the capstone project.

# Utilization of the Dataset

The datasets described will provide useful insights to Ford in the automobile industry. The researcher will use the datasets to generate descriptive and predictive analytics to assist Ford in assessing whether the company’s current product offerings align with the current and future states of the industry. The EPA dataset will generate useful predictive indicators on the change in production share of various automobile segments. Ford can compare the company’s current lineup to the predictive trend and determine whether their strategic goals align with how the market is trending into the future. The researcher will also use the EPA dataset to predict future miles per gallon and CO2 emissions for the different car segments and overall. This values and the future trends will assist Ford in determining whether their lineup is competitively viable in a market that is focusing more on environmental options (Høyer, 2008). Ford will also gain an understanding on the different technologies, such as transmission types being used in the marketplace, that are being deployed over time. The datasets were chosen to provide a trusted source of data that is continually updated on a yearly basis. The datasets are published from the federal government agencies of the EPA and the Department of Energy so the analysis conducted can be tested against future data entries into these datasets for continual reference for a company such as Ford. The market is evolving in the automobile sectors in areas such as the introduction of electric cars and whether this new introduction will replace standard gasoline cars, and if so, when this will occur in the future (Enyedi, 2018). These datasets and the application of data analytical tools will provide insights into these trends to allow Ford to remain as a top automobile manufacturer and create avenues of future research such as predicting the selling price of automobiles in future years (Idris et al., 2020).

# Objectives

The research conducted in this report is focused on providing the selected organization insightful and actionable analysis for industry knowledge. This focus is accomplished through objectives established by the researcher that will be explored through hypothesis testing in the research study. The primary objective is to determine the present and future states of the automobile industry in regards to gas mileage ratings of various vehicle types and the trend in the overall industry. The researcher will analyze the current data in the area of gas mileage ratings to ascertain the trend in the market in comparison to Ford’s current offering of products. The researcher also will determine the predictive future state of the market in regards to product offerings and expected gas mileage ratings. These predictions will be compared to current research on the topic and provide a framework for Ford Motor Company to conduct future analysis.­­

The secondary objective is to determine areas of future growth in the automobile industry. The primary objective serves to provide Ford Motor Company a model to assess competitive viability in the current and short-term future market, however, the secondary objective of the analysis is concerned with the long-term implications of the analysis. This includes the perceived trend of new technologies in the industry such as electric automobiles and various transmission types that do not have the historical data presently to generate a robust model. These future trends will be analyzed as part of the research to generate future actions that can be taken by the Ford Motor Company in the future market.

# Overview of the Study

The increasing concerns about environmental health and economic factors has applied pressure to many automobile manufacturers to introduce vehicles with higher gas mileage ratings to combat the emerging issues. These changes in the market demand have led to the purpose of the research conducted in the is report in assisting Ford Motor Company in navigating the market and current product trends. The Environmental Protection Agency is the regulatory body in the United States regarding the above issues, so the dataset used in the study is supplied by this organization as public data. The researcher will utilize this data to generate statistical models for the use of predicting future values for gas mileage ratings in the United States automobile industry to assess the rate of change and expected values in certain time periods. The study will also explore other factors related to this objective such as trends in new technologies that will impact the industry. The descriptive and predictive models will build upon current research in the area of environmental and economic concerns in the automobile industry to provide a holistic view of the future state of the industry, as well as, the areas of importance to the Ford Motor Company.

# Overview of Hypotheses

A hypothesis is defined as a “logical conjecture (hunch or educated guess) about the nature of relationships between two or more variables expressed in the form of a testable statement” (O’Leary, 2021, p. 50). These testable statements or hypotheses are going to frame the basis of the research project that the researcher is creating to assist the selected U.S. organization in demonstrating the analysis of a dataset using data analytical tools. The dataset(s) discussed in previous modules will provide the variables that will be tested to either prove or disprove the hypotheses. Toledo (2011) et al. assert that the research hypothesis can take multiple different forms including: inductive and deductive, directional and non-directional, and null and alternate hypotheses. The researcher will focus on the use of null and alternate hypotheses for the capstone project.

The hypotheses that will be used are derived from an understanding of the dataset in the analysis. The EPA automotive trends dataset provides multiple areas of variable analysis that would be beneficial to the Ford Motor Company (the capstone project organization). The hypotheses that will be tested through statistical testing will include the variation of the variables over time to determine the basis for future predictions based on variable values, the grouping of variables and values for comparisons, and overall comparisons of historical data and company data to support the research hypothesis described later in this report. The hypotheses will be stated with the null hypothesis stating that there is no change or distinct difference between the input and output variables and the alternate hypothesis stating that there is a statistical difference between the output and variables being tested in the hypothesis statement. For example, the hypothesis from the EPA dataset may state in the null hypothesis that “there is no statistically significant relationship between real-world miles per gallon and production share of vehicle types” while the alternate hypothesis would state that “there is a statistically significant relationship between real-world miles per gallon and production share of vehicle types”. These types of hypothesis testing will occur to determine the current and future states of the automobile industry to ascertain the factors impacting the products offered and characteristics of these products in the market. The main purpose of the analysis will be to determine the input variables that drive a change in real-world miles per gallon and determine the expected future values of this variable for the different regulatory/vehicle class types. The researcher will explore the statistical impact of the other variables in determining a model for predicting future real-world miles per gallon and the future of the production shares of the various vehicle types. These outputs will be compared to the current offerings of Ford Motor Company to compare the industry values with the organization’s current offerings which is data also contained in the EPA dataset. The analysis contains the model years for time forecasting methods to be applied, in addition to, variable testing to provide trend analysis which will be discussed within the research methods section of this report.

# Research Hypothesis

This report has thus far listed the objectives and overview of the intended study the researcher plans to conduct within the automobile industry for the Ford Motor Company. The research’s main focus is investigating the various factors that impact the real-world miles per gallon ratings and how these factors are changing over time to support the future state of the automobile industry. This research focus creates the basis for the research hypothesis that will be tested which can be stated as “What factors have statistically significant impact on the real-world miles per gallon values for automobiles?”. This emphasis will lead to testing the change in a variety of variables in the automaking process such as transmission types, vehicle classes offered, and the weight of vehicles. The research hypothesis will test to determine if these changes have significantly impacted the miles per gallon variable and to which variables Ford Motor Company should expend resources on to optimize the miles per gallon output of their vehicles in the competitive marketplace. This hypothesis will be tested using various descriptive and predictive statistical tests to either prove or disprove the null hypothesis which will be stated in the *Research Design* section of this report.

# Literature Review

The automobile industry has noticed aa changing environment, in relation to, the increase in the miles per gallon ratings and reduction in greenhouse emissions provided by the vehicles in the market. This change has been brought about by the different regulatory organizations in the United States that are requiring vehicle manufacturers to cooperate with environmental concerns. The Environmental Protection Agency (EPA) and the National Highway Traffic Safety Administration (NHTSA) are two of these regulatory agencies that have released documentation in regards to average fuel economy and greenhouse gas emission values for the automobile sector. These regulatory documents have provided a basis for the research conducted in this report, as well as, the literature and scholarly sources that will be highlighted in this section of the report that provided insights into the research project.

The main literature article used in support of the research conducted was “Vehicle efficiency and tractive work: rate of change for the past decade and accelerated progress required for U.S. fuel economy and CO2 regulations” by John Thomas published in *SAE International Journal of Fuels and Lubricants*. Thomas (2016) published research that investigated the EPA’s regulatory requirements for Corporate Average Fuel Economy (CAFÉ) and greenhouse gas (GHG) emission for the years of 2017 to 2025. The research provided insights into the predictions of future fuel mileage ratings and the rate of change of these values to meet the regulatory requirements by the year 2025. Thomas utilized historical data of conventional gas mileage vehicles from 2005 to the year 2015 as a basis for the predictions which is similar to the research conducted in this report with historical data being used from 1975 to 2020. Thomas researched the impact of a variety of variables on the gas mileage ratings such as transmission and powertrain types as the basis for the study. A similar approach was deployed by Wu and Liu (2011) in the article published in *Expert Systems with Applications* entitled “Development of a predictive system for car fuel consumption using an artificial neural network”. This article describes the process and need to build a predictive model for fuel consumption ratings due to regulatory and economic factors. The model proposed uses similar variables to this report’s model such as engine style, vehicle type, and transmission type etc. Wu and Liu found through their research that a statistically significant predictive model can be built using the proposed input variables and a neural network. This report utilizes a different model for future predictions but the basis of Wu and Liu’s model and assumptions provide reference for the research project. Factors were also referenced from the article “Mileage efficiency of cars” by Patankar et al. (2021) that described the characteristics taken into account when calculating fuel mileage ratings of an automobile.

The article entitled “The Macroeconomic Effects of 2017 through 2025 Federal Fuel Economy and Greenhouse Gas Emissions Standards by Carley et al. (2019) provides analysis on the economic factors that organizations can expect due to the change in miles per gallon and greenhouse emissions ratings in the United States. This source details the changing environment that the regulatory bodies will have on the automobile market and the cost of failing to meet regulatory requirements. Carley et al. agree with the research in this report that the rising values of gas mileage ratings will have an economic and market product offering impact on future vehicles in the automobile industry. This impact must be realized by automakers such as Ford Motor Company to maintain compliance and remain competitively viable especially given the changing environment on gasoline prices and employment impacts in the automaker industry. The future state of the automobile industry was referenced using the “Forecasting the trajectory of electric vehicle sales and the consequences for worldwide CO2 emissions” by Rietmann et al. (2020). Rietmann et al. provided research on the predictions of the future electric vehicle market, as well as, the impact the sales of these vehicles will have on the gas mileage and greenhouse gas emissions researched in this report. The logistical infrastructure concerns and production limitations were discussed and provided reference to this report’s future analysis from the data analytics performed on the EPA dataset.

The sources described in the literature review have provided references to models already built in the area of gas mileage predictions in the automobile industry. The researcher used these resources to gain understanding of the possible areas of expansion and the limitations that the developed predictive model can contain in the research project. The scholarly sources also provided knowledge into the current state of the industry and the avenues for future states that will be explored in the data analysis of the research project.

# Common Challenges in Security, Privacy, and Ethics of Data

The evolution of data collection and storage has created an environment rich with data availability in all forms. This change has been brought about in part to the datafication of many everyday tasks of the individual and the common need for convenience sought after by consumers. For example, in the retail industry consumers’ purchases are tracked, stored, and analyzed by organizations to determine what sales to publish or what products to market to which selection of consumers. This data is created through the various activities of the consumer from the purchasing action to searching on the website to the creation of an account with preferences chosen towards specific product offerings. These actions create a streamline of data that can be stored in more available data storage facilities and analyzed using modern models. This availability of data to a variety of organizations brings, with the information, concerns about the security, privacy, and ethics of user’s personal data.

## Security

Data security is defined by IBM as “the practice of protecting digital information from unauthorized access, corruption, or theft throughout its entire lifecycle” (“What is Data Security?”, 2021). The ability for an organization to protect the data within the databases that stores the data about consumers and the business is highly important for data analytics. There are common issues (especially in the area of big data) that challenge security the data in many companies. These issues include: Accidental Exposure, Phishing/Social Engineering Attacks, Insider Threats, Ransomware, etc. (Bekker, 2018).

The security concerns for data can be intentional or accidental depending on the type of breach. Accidental data breaches can occur due to accesses granted to employees without the need or training to properly handle sensitive information. Intentional data breaches are more difficult to mitigate for concerns such as phishing or ransomware. Phishing is a common tactic that allows outside users to gain access to data through malicious digital attacks on individuals in an organization while ransomware is malware that infects computers or servers’ systems to destroy data. These concerns are in tandem with the general security concerns of losing data via crashes in the data system either physically or digitally on the cloud (Gahi et al., 2016).

## Privacy

Data privacy is the concerned with the protection and handling of sensitive data that an organization may have on an individual (personal data) or company data. The Storage Networking Industry Association highlights that data privacy and data security are not interchangeable in stating that “data security protects data from compromise by external attackers and malicious insiders whereas data privacy governs how the data is collected, shared, and used” (“What is data privacy”, 2021). Data privacy largely is encompassed within organizations’ privacy policies that users are prompted with and agree to for using certain services. The main concern for users is how the data that is being collected and stored by the organization is being used and shared within the context of data privacy. The data security concerns, referred to previously in this report, monitors the malicious use of data but organizations must also comply with the privacy of data even while using it within a non-malicious context. Government regulations such as HIPPAA, General Data Protection Regulation (GDPR) in Europe, and the California Consumer Privacy Act (CCPA) assist organizations in assessing concerns with data privacy (Harding et al., 2019). The concerns are in areas of transparency, limitations on which data can be collected and storage limits, and data accuracy. Consumers expect that businesses be forthright with information on what data is being collected, how much data is collected and stored, and the ways the data is being used or shared with other organizations. These concerns must be asked of any data analytical project to ensure compliance with regulations and the goodwill of the consumer whose data is being utilized for the analysis.

## Ethics

The idea of data ethics can be described as “encompassing the moral obligations of gathering, protecting, and using personally identifiable information and how it affects individuals” (Cote, 2021). Davis and Patterson (2012) further describe data ethics to include four elements: Identity, Privacy, Reputation, and Ownership. Identity refers to the ability of data to provide identifiable characteristics that can be traced to an individual person which is a concern for autonomy that cannot be exploited by an organization. This identity concern can impact the reputation of an individual depending on how the data collected is used and in what context by an organization. This is especially true in the areas of privacy and ownership. Data ownership is the concern that there is no defined point in which an organization or an individual can assert that the data collected is the property of one of the entities. Organizations have an obligation to treat data ethically to protect the consumer or user from any negative effects due to the analysis of data (O’Leary, 2016).

# Research Design

The research design will include details on the methodology the researcher deployed as the framework for the data analysis. The methods and tools used throughout the research analysis are discussed, as well as, the models produced in support of the objectives to answer the research hypothesis. This section also describes the limitations of the research and the ethical concerns that were discovered and will be considered in relation to the project.

## Methodology

The methodology used for the research project leveraged statistical tools and methods to test the relationship of variables within the selected public dataset. The research conducted for the research project included using derived or compiled data from historical datasets. Derived data uses existing data to generate new information through various processes such as model development (“Data Module #1: What is research data?”, 2021). The existing data that was used included the EPA dataset and complimentary dataset such as the global electric car market and top automobile manufacturers datasets provided by the U.S. Department of Energy. These datasets were inputted into the models for descriptive and predictive statistics. The descriptive statistical tests primarily focused on distribution analysis of the variables using histograms and measures of central tendency (Kaur et al., 2018). The methods varied for the different types of variables within the dataset such as numerical or categorical variables.

The descriptive statistical tests provided inputs into the creation of the predictive model which tested the research project’s null hypothesis stated as “there is no statistically significant relationship between the variables in the EPA dataset and real-world miles per gallon values” with the alternate hypothesis stated as “there is a statistically significant relationship between the variables in the EPA dataset and real-world miles per gallon values”. The predictive models created were used to either reject or accept the research project’s hypothesis using multi-variable linear and time-series forecasting models which are described in the *Methods* section of the research design.

## Methods, Tool, and Techniques

The methods used include the tools and techniques deployed to gain insights from the dataset in relation to the research hypothesis. The methodology described the use of descriptive statistical tests which were conducted with the Python coding language. The researcher used the Jupyter notebook distribution to create descriptive models on distributions, measures of central tendency, and the overall nature of the dataset. The Pandas, Numpy, and Matplotlib library distributions in Python were leveraged to numerically and visually examine the variables with the book *Python for Data Analysis* by Wes McKinney as the primary reference. The analysis included insights into the viability of the dataset and the expected relationships that could be expected in predictive model building in the future state.

The predictive models that are used for the research project utilized the R programming language in the R Studio environment. The models leveraged the linear model (“lm”) function in R for multi-variable linear regression where real-world miles per gallon was the output and stepwise regression was used for the selection criteria. The model was evaluated using the AIC values and prediction coefficients. A model was also built using the ARIMA (autoregressive integrated moving average) function in R to compare stepwise regression and times series ARIMA models for predicting future values of the output variable. The guides produced by Singh (2019) and on Statology.org (2021) provided reference in creating these models. These models are the basis for rejecting or accepting the null hypothesis for the research project and provide insights into the input variables, that are within Ford Motor Company’s control, that significantly impact the real-world miles per gallon variable.

## Limitations

The limitations of the research project are found within the availability of data. The EPA dataset provides useful information for the descriptive and predictive model building for the selected organization for understanding the market, however, the data needed to predict the expansion of new technologies is limited at the present time. These limitations in data and not in research allows only for the researcher to speculate as to the future state of these technologies and how they will impact the product market in a forward-leaning approach. The models presented are also limited to the variables presented in the EPA dataset and could provide a more rigorous model if internal Ford Motor Company data was inputted to the model.

## Ethical Considerations

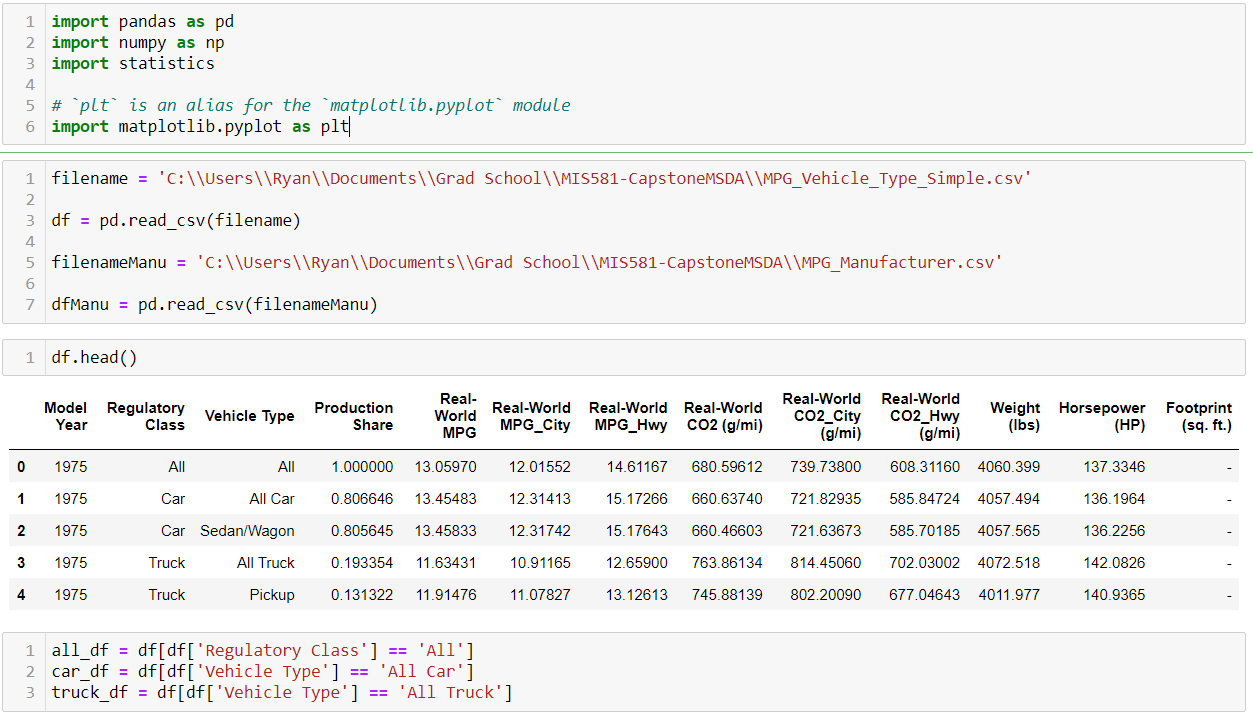
The utilization of data analytics brings forth the requirement to consider ethical concerns in the dataset and its applications. Davis and Patterson (2012) describe data ethics to include four elements: Identity, Privacy, Reputation, and Ownership. Identity refers to the ability of data to provide identifiable characteristics that can be traced to an individual person which is a concern for autonomy that cannot be exploited by an organization. This identity concern can impact the reputation of an individual depending on how the data collected is used and in what context by an organization. This is especially true in the areas of privacy and ownership. Data ownership is the concern that there is no defined point in which an organization or an individual can assert that the data collected is the property of one of the entities. Organizations have an obligation to treat data ethically to protect the consumer or user from any negative effects due to the analysis of data (O’Leary, 2016). The researcher placed these concerns under consideration with the use of the EPA dataset. The concerns are alleviated in knowledge of the EPA trends report dataset containing publicly available federal data that is aggregated to mitigate ethical concern risks to the individual. The researcher utilized the aggregated data which mitigates the risks of identity, reputation, and privacy as the data cannot be linked to an individual or organization. The dataset is also a combination of regulatory data which corresponds to the ownership concern as the United States federal government has requirements to be in ownership of the data in the dataset.

The data analysis and findings of the research project will be utilized to fulfill the objectives of the research project and provide insights into the research hypothesis. The *Findings* section of this report will detail the graphical and statistical outputs of the statistical tests described in the research design, in addition to, the results of the data analysis of the dataset. The researcher will also provide a conclusion and future recommendations for areas of future research that the selected organization will want to explore further with the provided models.

# Data Analysis and Findings

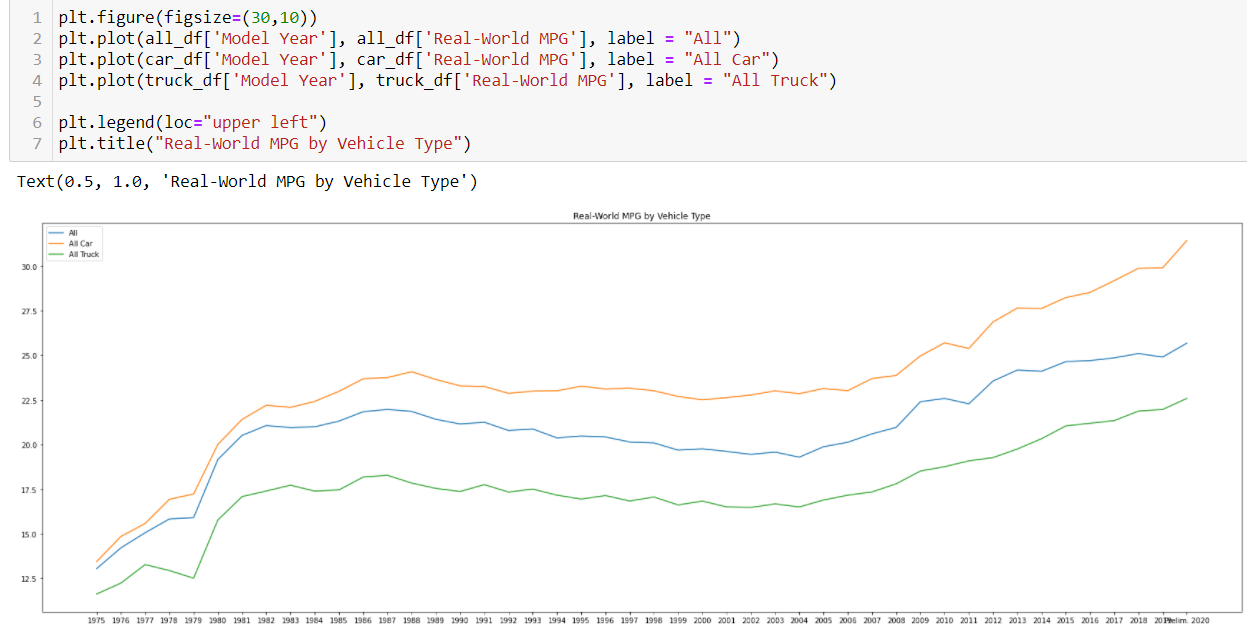
The research project design provided the framework for the data analysis that will be presented in this section of the report. Code statement and outcome screenshots will be provided in the report and the full dataset with model code can be found on the researcher’s GitHub: <https://github.com/rymey21/CapstoneProjectMSDA.git> . The first step of the data analysis included the exploration and generation of descriptive statistics of the EPA dataset in Python Jupyter Notebooks. The researcher imported the Numpy, Pandas, Statistics, and Matplotlib libraries to utilize in the analysis. The EPA dataset was filtered to generate data frames in the Python environment for the different regulatory classes of “All”, “All Car”, and “All Truck” so that different visuals could be created depending on the use case. The library and dataset importing, as well as, the filtering of the dataset can be seen in *Figure 1*.

**Figure 1**

Code statements for importing and filtering the EPA dataset.

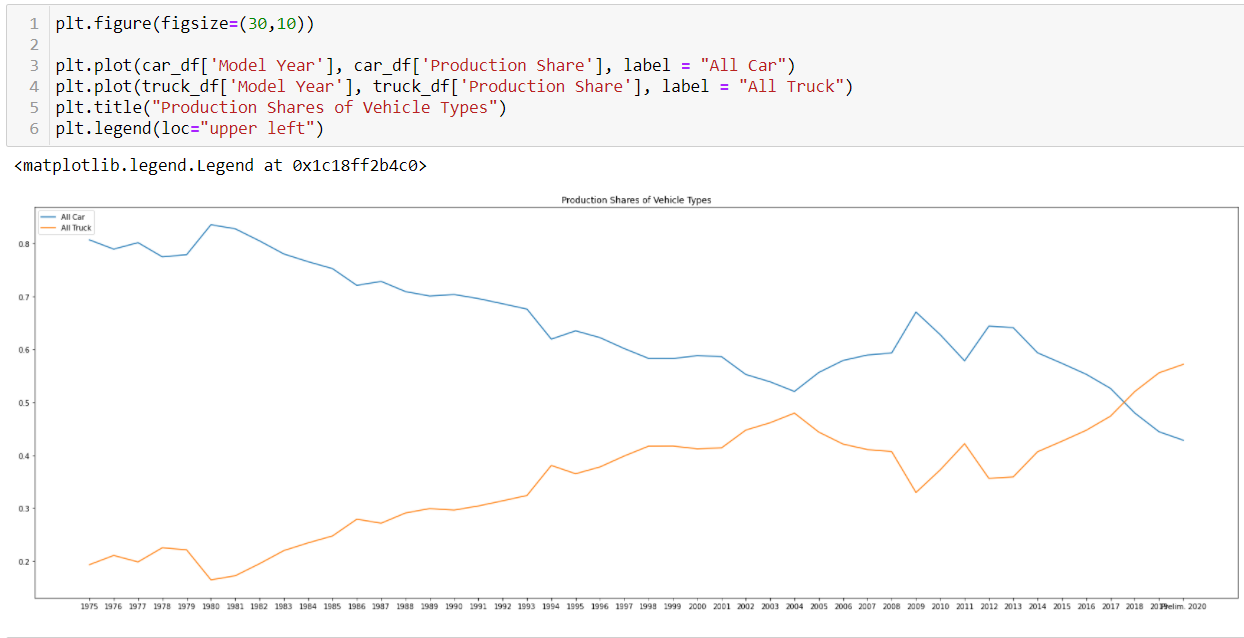
The filtered datasets were used for the exploratory analysis. The first visual built was a line chart of the real-world MPG values over model year for the three regulatory classes which is shown in *Figure 2*. The figure shows that there has been an increase in real-world MPG ratings over time with the highest increase in the car regulatory class and the lowest being in the truck regulatory class.

**Figure 2**

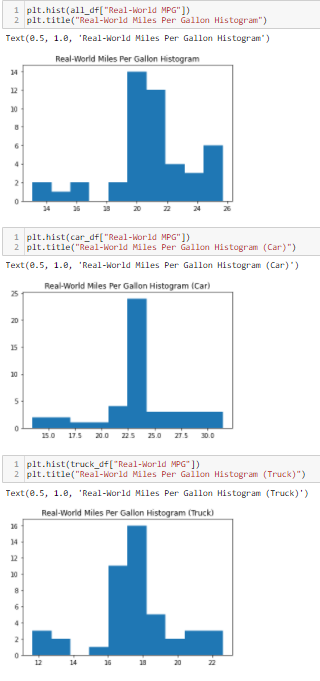
Line chart of Real-World MPG values per regulatory class.

The differences in Real-World MPG values over time by the regulatory classes lead the researcher to continue investigating the relationship between the different regulatory classes and miles per gallon values. This analysis included analyzing the production share of vehicle types for the car and truck regulatory classes over time, in addition, to performing descriptive statistical tests on the distribution of Real-World MPG values in the dataset. The line chart in *Figure 3* shows that recently (the year 2018) the demand for truck surpasses the demand for cars in the United States which means that the higher demand regulatory class has the lower miles per gallon ratings. The distributions of the Real-World MPG values in *Figure 4* highlights that the values for miles per gallon in all three regulatory filters can be assume to be normally distributed with slight skewness. This distribution allows for further predictive analytical models to be built that requires a normally distributed response variable.

**Figure 3**

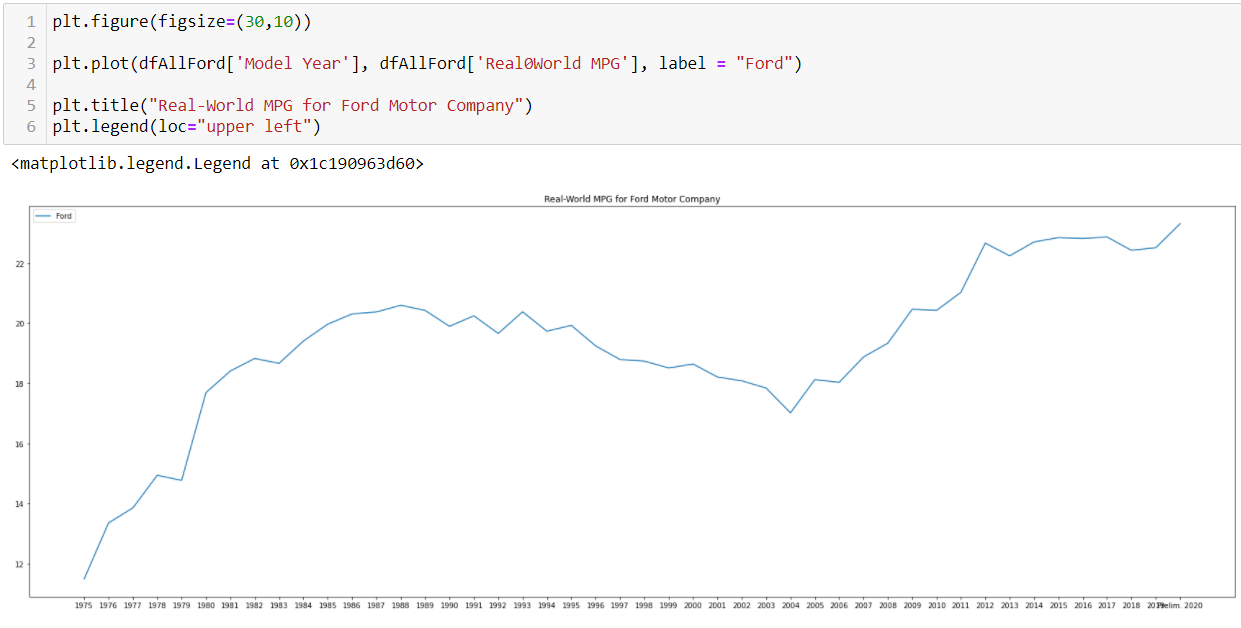
Production share of regulatory classes over time in the United States.

**Figure 4**

Histogram distributions of the Real-World MPG variable for the regulatory classes.

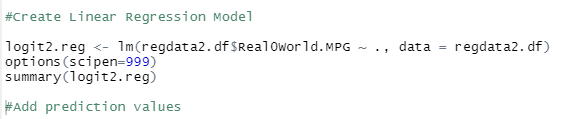
The descriptive statistical analysis proved that the variables in the dataset had potential to provide statistically significant relationships to build a predictive model for miles per gallon value predictions. The researcher would use this analysis for the creation of the analytical models in the R programming language. The exploratory data analysis also yielded some more information that would be insightful for the Ford Motor Company. *Figure 5* displays a line chart for the Real-World MPG values of the Ford Motor Company for the same date range as the regulatory required data presented previously in the report. These values can be used as a benchmark to compare the current abilities of Ford in producing fuel-efficient vehicles.

**Figure 5**

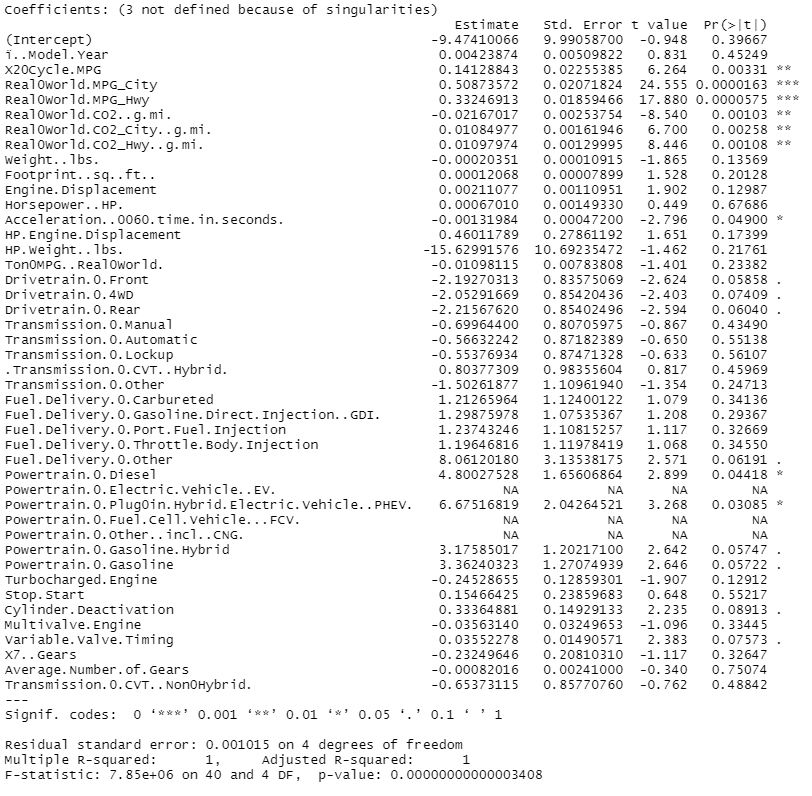
Ford Motor Company Real-World MPG values over time.

The predictive models that were built by the researcher include the regression and times series forecasting (ARIMA) models described in the *Research Design* portion of this report. The dataset was partitioned in the same slices as was conducted for the descriptive statistical tests with the “All” regulatory and vehicle types being used for the predictive model analysis. The variables were coerced to be numeric as the original dataset had these numeric variables as class variables and the non-numeric variables were removed from the model due to the nature of the regression and ARIMA models requiring numeric input variables. The “lm” function for standard linear regression model was used in R Studio using the variables in EPA dataset to generate a predictive model with Real-World MPG is the response/predicted variables. The code statements and output of the model are shown in *Figures 6-7*.

**Figure 6**

Code statements for the linear regression model in R Studio.

**Figure 7**

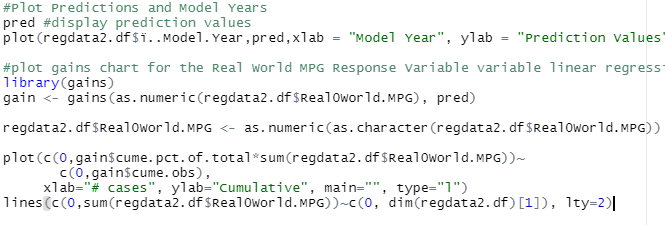
Linear regression model output.

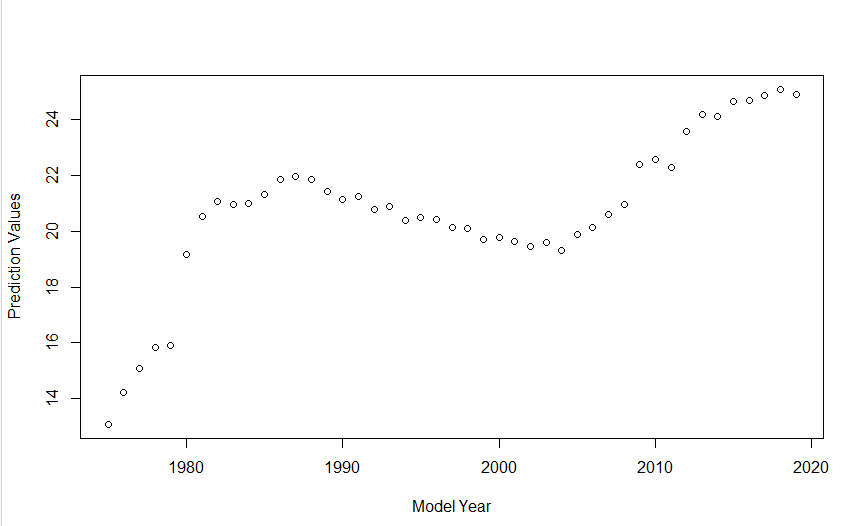
The linear regression model displays that the overall model is statistically significant with a p-value less than 0.05 for a 95% confidence interval. The output also shows that variables within the model are have a statistically significant impact on the average real-world MPG values in the dataset such as acceleration, CO2, and powertrain options with p-values below or close to below the 0.05 threshold. This is basis to reject the research null hypothesis in favor of the alternate hypothesis which states that there are statistically significant relationships between vehicle variables in the EPA dataset to the real-world MPG values. The rejection of the null hypothesis allows this model to be used for predictions of future real-world MPG values by Ford Motor Company. The linear regression model is utilized by creating a linear equation where the estimate values are the effect on the response variable. For example, the diesel powertrain variable would have an increase in real-world MPG values by 4.8 MPG for every percent increase in this powertrain. This model also highlights the variables or vehicle attributes that have the largest impact and the type of impact on MPG ratings thus the areas Ford Motor Company should focus on going forward in designing products. The dataset was also filtered to provide a basis to build models on the different vehicle and regulatory types if Ford Motor Company desired to complete further analysis on a certain product segment.

The researcher used the regression model to generate predictive values for the dataset and compare the model’s performance to a naïve approach using a gains chart, as well as, a chart of the prediction values. The gains chart demonstrate that the regression model outperforms the naïve approach and is a better predictive model for computing real-world MPG values. The code statements and associated charts are shown in *Figures 8-10*. The residuals plot is also shown in *Figure 11-12* to demonstrate that the model has evenly distributed residuals and therefore is an accurate model.

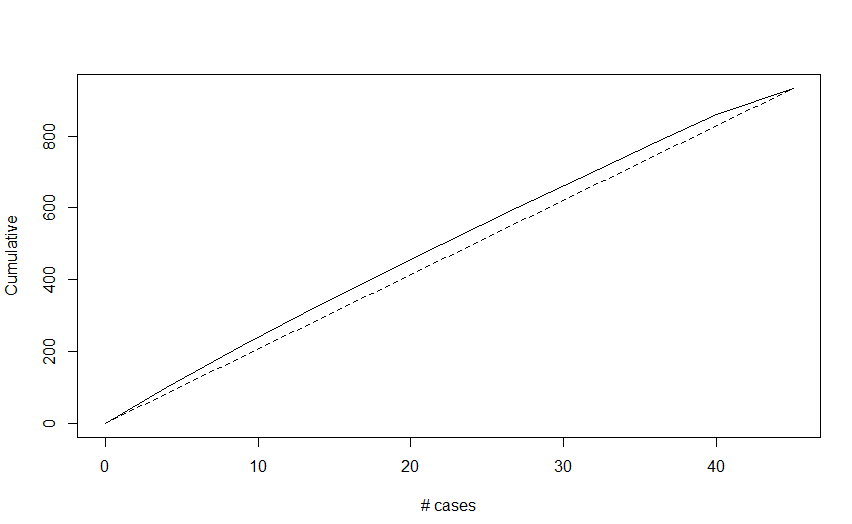
**Figure 8**

Code statements for the model prediction value and gains/lift chart.

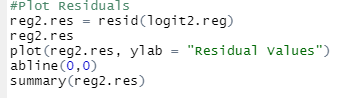
**Figure 9**

Prediction values chart.

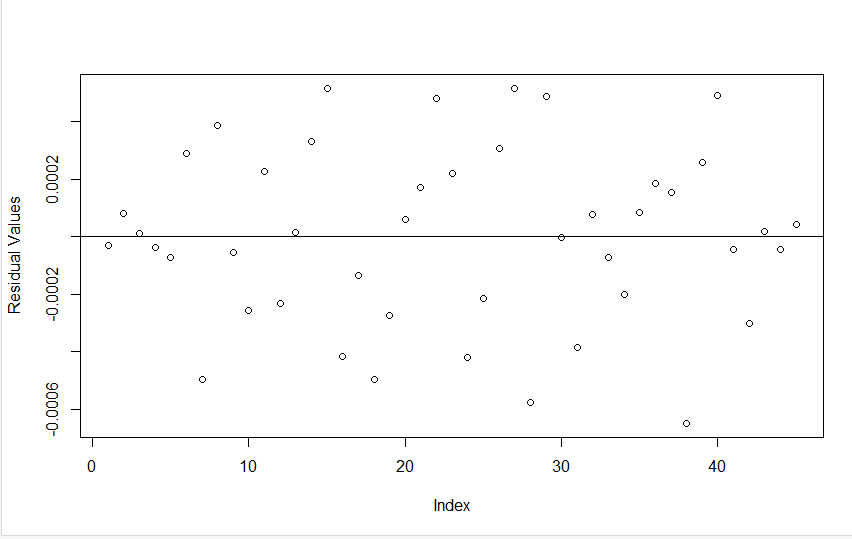
**Figure 10**

Gains chart output for the regression model.

**Figure 11**

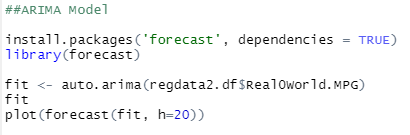
Code statements to generate residuals and residual plot.

**Figure 12**

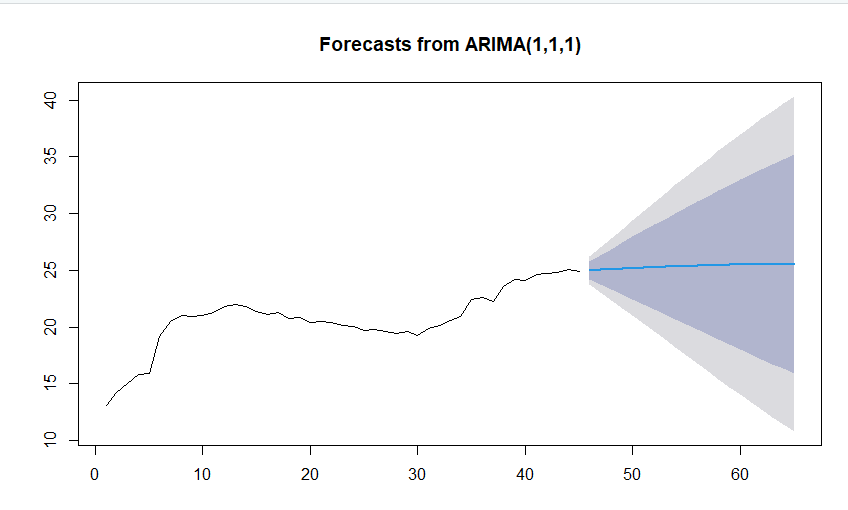
Residual plot of the regression model to demonstrate evenly spread residuals.

The researcher also built an autoregressive integrated moving average (ARIMA) time series forecasting model for the same dataset to provide a secondary model for predicting values using the “forecast” library in R. The ARIMA model’s solution was statistically significant and provides similar predicted values to the regression model. The code statements for the ARIMA model and the ARIMA graphical output are shown in *Figures 13-14*. The graphical output of the ARIMA model displays that the blue line is the area of highest prediction for the model with the ability of the model to spread to the dark gray and light gray area over a longer time period.

**Figure 13**

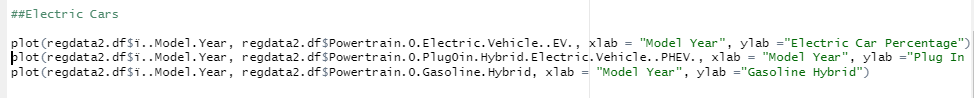
Code statements for the ARIMA model.

**Figure 14**

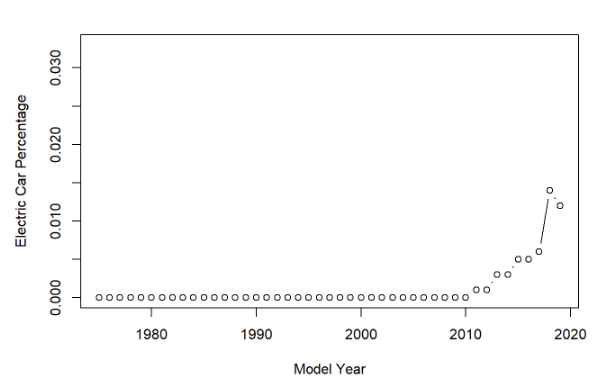
Graphical output of the ARIMA model.

The output of both the regression model and ARIMA justify the research hypothesis by proving statistically significant relationships between variables in the EPA dataset and real-world MPG values. This outcome highlighted a change in the automobile industry that was furthered analyzed in the area of electric and hybrid vehicles. The same dataset was utilized to generate charts of the production share of electric and hybrid vehicles with the code and outputs shown in *Figures 15-18*.

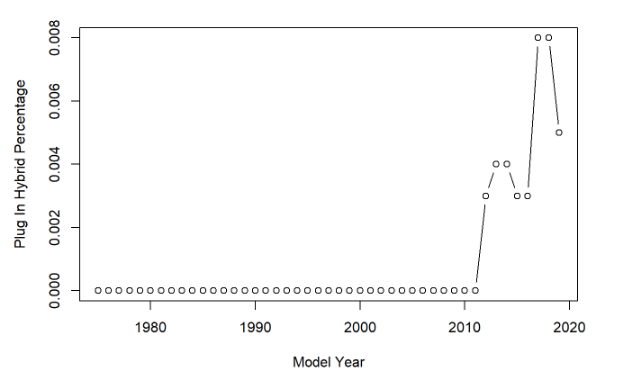
**Figure 15**

Code statements for the electric and hybrid visual charts.

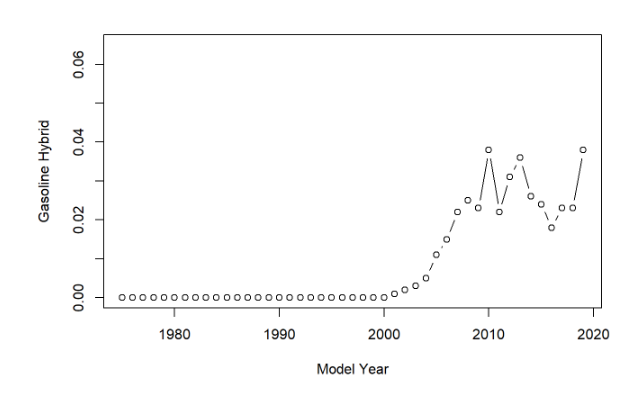
**Figure 16**

Electric vehicle production share chart.

**Figure 17**

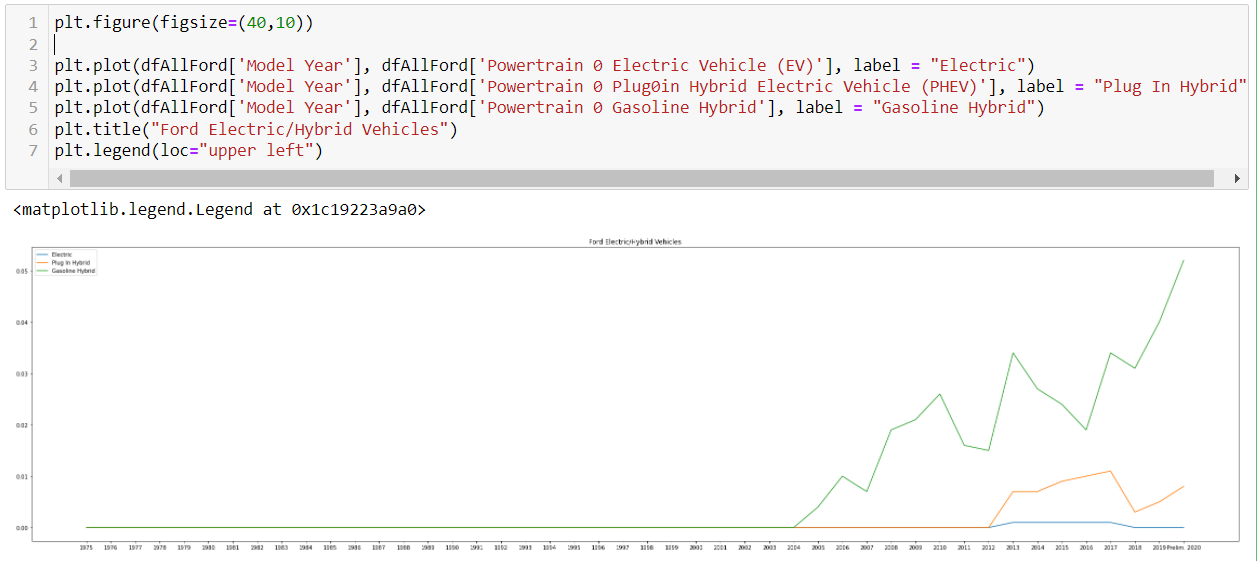
Plug-In hybrid production share chart.

**Figure 18**

Gasoline hybrid production share chart.

These visuals demonstrate that there is a rising demand in producing vehicles with alternate powertrain types such as electric and hybrid vehicles. The predictive models showed that these variables have an impact on increasing real-world miles per gallon values but the current demand does not match the demand for more traditional powertrain types such as gasoline or diesel. The researcher believes that these alternate powertrain types are an area of future research for Ford Motor Company. In comparison to the company’s current lineup from the EPA dataset, the organization is following track with the emerging trends as shown in *Figure 19*.

**Figure 19**

Ford Motor Company’s electric and hybrid production values.

# Conclusion

This report has provided research into the real-world miles per gallon values, trends, and future of the automobile industry. The information and data analysis in this research report can be utilized by Ford Motor Company in predicting future values in the area of miles-per gallon ratings to generate a competitive advantage for the various products in the company’s lineup. The objectives of the project were accomplished, as well as, disproving the research null hypothesis through descriptive and predictive statistical tests and modelling. The linear regression model and ARIMA model provided a strong framework in demonstrating how a similar dataset can be analyzed by Ford Motor Company in the future. The descriptive statistical tests also demonstrate how data can be explored and altered for models and analysis. The real-world miles per gallon values are climbing in response to a variety of regulatory and economic factors that will shape the future of the automobile industry as it continues to evolve with new technologies. These emerging factors can be additional variables added within the constructed predicted model to generate insights into the impact that Ford Motor Company can expect the changing environment to have in the industry and to the company’s products.

## Recommendations

The researcher recommends that the models and analysis provided in the report be used by the Ford Motor Company for understanding the changing environment in the automobile industry. The models can be used for predicting future values of various vehicle types to provide a basis for providing competitively viable products for in demand vehicle variables such as real-world MPG values. The analysis also explored areas of future research such as the climb in demand for alternate powertrain vehicles such as electric and hybrid options. Ford Motor Company can input internal data for the various vehicles offered to fine-tune the model to actual data that is confidential to the organization. This report can also be used for a framework for different data analyzes and the required steps or considerations when performing data analysis projects. The research project provided in this report is inclusive of the data analytical lifecycle and provided justification for a statistically significant predictive model that can be leveraged for present and future product specification generation in a competitive market.

# References

*A complete guide to stepwise regression in R.* Statology.(2021, August 25).Retrieved October 30, 2021, from https://www.statology.org/stepwise-regression-r/.

A&E Television Networks. (2009, November 13). *Ford Motor Company Incorporated*. History.com. Retrieved October 15, 2021, from https://www.history.com/this-day-in-history/ford-motor-company-incorporated.

Carley, S., Zirogiannis, N., Duncan, D., Siddiki, S., & Graham, J. D. (2019). The Macroeconomic Effects of 2017 Through 2025 Federal Fuel Economy and Greenhouse Gas Emissions Standards. *Journal of Policy Analysis & Management*, *38*(3), 732–763. https://doi.org/10.1002/pam.22132.

Cote, C. (2021, March 16). *5 principles of data ethics for business*. Business Insights - Blog. Retrieved November 6, 2021, from https://online.hbs.edu/blog/post/data-ethics.

*Data module #1: What is research data?: Types of research data*. DeWitt Wallace Library . (2021, March 9). Retrieved October 29, 2021, from https://libguides.macalester.edu/c.php?g=527786%2F&p=3608643.

Davies, T. M. (2016). *The book of r: A first course in programming and Statistics*. No starch Press.

Davis, K., & Patterson, D. (2012). *Ethics of Big Data: Balancing risk and innovation*. O'Reilly.

Environmental Protection Agency. (2021, January). *Automotive Trends Report*. EPA. Retrieved October 16, 2021, from https://www.epa.gov/automotive-trends/explore-automotive-trends-data.

Enyedi, S. (2018, May). Electric cars—Challenges and trends. In *2018 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)* (pp. 1-8). IEEE.

Evans, C., Cheah, L., Bandivadekar, A., & Heywood, J. (2009). Getting more miles per gallon. *Issues in Science and Technology*, *25*(2), 71-80.

*Ford Again Leads Automakers in U.S. Assembly, Hourly Jobs, Underscoring Commitment to American Economy*. Ford Media Center. (2021, March 19). Retrieved October 16, 2021, from https://media.ford.com/content/fordmedia/fna/us/en/news/2021/03/19/ford-leads-automakers-us-assembly-vehicle-sales.html.

*Ford's impact in the US*. Ford Corporate. (2021). Retrieved October 16, 2021, from https://corporate.ford.com/social-impact/economic-impact.html.

Fortune. (2021, August 3). *Ford Motor: 2021 Fortune 500*. Fortune. Retrieved October 16, 2021, from https://fortune.com/company/ford-motor/fortune500/.

Gahi, Y., Guennoun, M., & Mouftah, H. T. (2016, June). Big data analytics: Security and privacy challenges. In *2016 IEEE Symposium on Computers and Communication (ISCC)* (pp. 952-957). IEEE.

Harding, E. L., Vanto, J. J., Clark, R., Hannah Ji, L., & Ainsworth, S. C. (2019). Understanding the scope and impact of the California Consumer Privacy Act of 2018. *Journal of Data Protection & Privacy*, *2*(3), 234-253.

Høyer, K. G. (2008). The history of alternative fuels in transportation: The case of electric and hybrid cars. *Utilities Policy*, *16*(2), 63-71.

Idris, N. O., Achban, A., Utiarahman, S. A., Karim, J., & Pontoiyo, F. (2020, November). Predicting the Selling Price of Cars Using Business Intelligence with the Feed-forward.

Kaur, P., Stoltzfus, J., & Yellapu, V. (2018). Descriptive statistics. *International Journal of Academic Medicine*, *4*(1), 60.

McKinney, W. (2017). *Python for Data Analysis, 2nd Edition*. O'Reilly Media, Inc.

*Locations*. Ford Corporate. (2021). Retrieved October 16, 2021, from https://corporate.ford.com/operations/locations.html.

O'Leary, D. E. (2016). Ethics for big data and analytics. *IEEE Intelligent Systems*, *31*(4), 81-84.

O’Leary, Z. (2021). *The essential guide to doing your research project.* (4th ed.). SAGE Publishing.

Patankar, N. A., Lin, J., & Patankar, T. N. (2021). Mileage efficiency of cars. *Cleaner Engineering and Technology*, *4*, 100240.

Rietmann, N., Hügler, B., & Lieven, T. (2020). Forecasting the trajectory of electric vehicle sales and the consequences for worldwide CO2 emissions. *Journal of Cleaner Production*, *261*, 121038.

Singh, D. (2019, July 12). *Time Series Forecasting Using R*. Pluralsight. Retrieved October 29, 2021, from https://www.pluralsight.com/guides/time-series-forecasting-using-r.

Sood, S. K. (2012). A combined approach to ensure data security in cloud computing. *Journal of Network and Computer Applications*, *35*(6), 1831-1838.

*The global electric car market*. Knoema. (2020, February 4). Retrieved October 17, 2021, from https://knoema.com/infographics/ffgfqj/the-global-electric-car-market.

Thomas, J. (2016). Vehicle efficiency and tractive work: rate of change for the past decade and accelerated progress required for U.S. fuel economy and CO2 regulations. *SAE International Journal of Fuels and Lubricants*, *9*(1), 290+. https://link.gale.com/apps/doc/A478973873/AONE?u=colstglobal&sid=bookmark-AONE&xid=aeeaf302

*Timeline - History, Chronology - Ford Rouge Factory Tour*. The Henry Ford. (2021). Retrieved October 15, 2021, from https://www.thehenryford.org/visit/ford-rouge-factory-tour/history-and-timeline/timeline/.

Toledo, A. H., Flikkema, R., & Toledo-Pereyra, L. H. (2011). Developing the research hypothesis. *Journal of Investigative Surgery*, *24*(5), 191-194.

*Top vehicle manufacturers in the US market, 1961-2016*. Knoema. (2020, May 21). Retrieved October 16, 2021, from https://knoema.com/infographics/floslle/top-vehicle-manufacturers-in-the-us-market-1961-2016.

*What is data privacy?* SNIA. (2021). Retrieved November 6, 2021, from https://www.snia.org/education/what-is-data-privacy.

*What is Data Security?* What is Data Security? Data Security Definition and Overview. (2021). Retrieved November 5, 2021, from https://www.ibm.com/topics/data-security.

Wee, H. M., & Wu, S. (2009). Lean supply chain and its effect on product cost and quality: a case study on Ford Motor Company. *Supply Chain Management: An International Journal*.

Wu, J.-D., & Liu, J.-C. (2011). Development of a predictive system for car fuel consumption using an artificial neural network. *Expert Systems with Applications*, *38*(5), 4967–4971. https://doi.org/10.1016/j.eswa.2010.09.155