Improving Automotive Battery Sales Forecast

by

Vinod Bulusu
Master of Business Administration, IE Business School, 2015
Master of Science, Chemical Engineering, University of New Hampshire, 2006

and

Haekyun Kim
Bachelor of Science, Mechanical Engineering, SungKyunKwan University, 2008

SUBMITTED TO THE ENGINEERING SYSTEMS DIVISION
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ENGINEERING IN LOGISTICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUNE 2015

© 2015 Vinod Bulusu and Haekyun Kim. All rights reserved.

The authors hereby grant to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Signature of Author..........................
Master of Engineering in Logistics Program, Engineering Systems Division
May 8, 2015

Signature of Author..........................
Master of Engineering in Logistics Program, Engineering Systems Division
May 8, 2015

Certified by...........................................
Dr. Roberto Perez-Franco
Research Associate, Center for Transportation and Logistics
Thesis Supervisor

Accepted by.................................
Dr. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering
Improving Automotive Battery Sales Forecast

by

Vinod Bulusu

and

Haekyun Kim

Submitted to the Engineering Systems Division
on May 8, 2015 in Partial Fulfillment of the
Requirements for the Degree of Master of Engineering in Logistics

Abstract

Improvement in sales forecasting allows firms not only to respond quickly to customers’ needs but also to reduce inventory costs, ultimately increasing their profits. Sales forecasts have been studied extensively to improve their accuracy in many different fields. However, for automotive batteries, it is very difficult to develop a highly accurate forecast model because many variables need to be considered and their correlations are complex. Additionally, current sales forecasts are derived from historical data and thus do not include any other causal factor analysis.

In this study we applied causal factor analysis to determine how the forecast accuracy could be improved. We focused on understanding the relationship between temperature and sales. Using regression modelling, we found that there is a quadratic relationship between temperature and battery sales. We validated the model by comparing the actual and predicted sales for various geographies and times. We concluded that the model is more robust for predicting sales across various times than through various geographies.

Thesis Supervisor: Dr. Roberto Perez-Franco
Title: Research Associate, Center for Transportation and Logistics
Acknowledgements

This effort is dedicated to my wife, Madhuri......thanks for being there always

I gratefully acknowledge the Office of the Dean for Graduate Education and O'Biren family for the generous fellowship throughout the program. I would also like to thank our thesis sponsor for letting us tap onto their knowledge to enable us to complete our thesis. I express my gratitude to Dr. Roberto Perez-Franco for his continuous support and encouragement even during challenging times. I would also thank Haekyun Kim who spent numerous nights working on this thesis for his positive energy and flexibility. I owe a huge debt to my wife, Madhuri and sons Advaita and Atharva for the time not spent with them.

I wish to express my sincere thanks to my thesis sponsor for providing this great opportunity. I am also grateful to Dr. Roberto Perez-Franco. I am extremely thankful and indebted to him for sharing expertise, and sincere and valuable guidance and encouragement extended to me. I take this opportunity to express gratitude to all of the Department faculty members for their help and support. I also thank my wife Eunjung for the unceasing encouragement, support and attention. I am also grateful to my thesis partner Vinod Bulusu who supported me through this venture.
Table of Contents

Abstract ........................................................................................................................................... 2
Acknowledgement ......................................................................................................................... 3
Figures ............................................................................................................................................. 5
Tables .............................................................................................................................................. 7
1. Introduction ................................................................................................................................ 8
2. Literature Review ....................................................................................................................... 9
   2.1 Models to predict the age of lead-acid battery ............................................................... 10
   2.2 Connecting Point of Sale (POS) to Forecasting ............................................................. 14
   2.3 Conclusion: The need for a multivariate model of POS data ....................................... 15
3. Methodology, Data and Analysis ............................................................................................. 17
   3.1 Overall Methodology ....................................................................................................... 17
   3.2 Data Collection ................................................................................................................ 17
   3.3 Modeling (data from 2010-2014) .................................................................................. 34
   3.4 Modeling validation ......................................................................................................... 39
4. Validation of the approach ......................................................................................................... 42
   4.1 Model G diagnostics ....................................................................................................... 44
   4.2 Model T diagnostics ....................................................................................................... 47
   4.3 Insights on the validity of the approach ....................................................................... 49
5. Conclusion and Future Work ....................................................................................................... 53
References .................................................................................................................................... 54
Figures

Figure 3-1: Process flow................................................................. 17
Figure 3-2: POS data composition .................................................. 19
Figure 3-3: SKU sales by time......................................................... 21
Figure 3-4: Zip code of sales .......................................................... 22
Figure 3-5: Top 10 sales by SKU .................................................... 23
Figure 3-6. Geographical sales of SKU 65 ....................................... 24
Figure 3-7: Aggregated sales in Boston area .................................... 25
Figure 3-8: Sales of 5 cities............................................................. 26
Figure 3-9: Temperature profiles of 25 stations in Boston area ............ 28
Figure 3-10: Average weekly temperature of the entire Boston area ... 28
Figure 3-11. Location of regions removed........................................ 29
Figure 3-12: 5 Stations not applicable to temperature aggregation .......... 30
Figure 3-13: Average weekly temperature of the entire LA area .......... 30
Figure 3-14: Temperature profiles of 10 stations in Houston area .......... 31
Figure 3-15: Average weekly temperature of the entire Houston area .... 31
Figure 3-16: Temperature profiles of 15 stations in DC area ............... 32
Figure 3-17: Average weekly temperature of the entire DC area .......... 32
Figure 3-18: Temperature profiles of 10 stations in Chicago area .......... 33
Figure 3-19: Average weekly temperature of the entire Chicago area .... 33
Figure 3-20. Diagnostics of Model 1 .............................................. 35
Figure 3-21. Diagnostics of Model 2 ................................................................. 36
Figure 3-22. Diagnostics of Model 3 ................................................................. 37
Figure 3-23. Diagnostics of Model 4 ................................................................. 38
Figure 3-24. Diagnostics of Model 5 ................................................................. 38
Figure 3-25. Diagnostics of Model 6 ................................................................. 39

Figure 4-1: Model diagnostics for Model G: $R^2$ ........................................... 44
Figure 4-2: Pareto Plot for Model G ................................................................. 45
Figure 4-3: Parameter Estimates for Model G ............................................... 45
Figure 4-4: Prediction Expression for Model G ............................................. 46

Figure 4-5: Model diagnostics for Model T: $R^2$ .......................................... 47
Figure 4-6: Pareto Plot for Model T ................................................................. 48
Figure 4-7: Prediction Expression for Model T ............................................. 48

Figure 4-8: Model validation for Boston (Model G) ....................................... 49
Figure 4-9: Model validation for Washington D.C. (Model G) ...................... 50
Figure 4-10: Model validation for Year 2012 (Model T) ............................... 50
Figure 4-11: Model validation for Boston (Model G) based on change .......... 52
Figure 4-12: Model validation for Washington D.C. (Model G) based on change .......... 52
Tables

Table 3-1: Sell-in & Sell-out data features........................................................................................................ 18
Table 3-2: Normalized sales............................................................................................................................ 27
Table 4-1: Validation of models....................................................................................................................... 42
Table 4-2: Minimum and Maximum Temperatures across the cities......................................................... 43
Table 4-3: Minimum and Maximum Temperatures across the years ....................................................... 43
1. Introduction

Our thesis sponsor is a global diversified technology and industrial leader serving customers in more than 150 countries. Especially, they are the global leader in lead-acid automotive batteries and advanced batteries for start-stop, hybrid and electric vehicles. Their market share almost reached 40% (2013) in the US. Our thesis sponsor sells batteries through major automotive service retailers such as AutoZone and traditional supermarkets such as Walmart.

In 2013, our sponsor company saw a phenomenal increase in automotive battery sales. Since the ramp up to production is a long process, the company could not meet demand. As a result, the company lost sales and also lost the opportunity to increase market share.

Because of this experience, our sponsor company is aware of the risks brought by the variability in demand and of the importance of forecasts. Their current forecasting model is based solely on historical sales data and does not include other variables which could influence battery failure and thus sales of new batteries. Since, our thesis sponsor’s mainly deals in after-market replacement battery sales, most of these sales occur due to a battery failure. Hence, battery failures correspond to battery sales. To prevent such a problem of lost sales and to meet unexpected market demand in a timely manner, a good forecast is highly desirable.

In this thesis, we will propose a methodology to improve sales forecast for our sponsor. Several previous studies have suggested that temperature has an impact on the failure rate of batteries. Therefore, in this thesis we explore the link between temperature and sales in the aftermarket battery. In the following chapters, we present our literature review, methodology, results, discussion, and conclusion of our thesis.
2. Literature Review

Many researchers such as Ruetschi (2004), Doerffel and Sharkh (2006), and Sauer and Wenzl (2007) have performed experimental and computational studies of the factors determining battery life. More recently, Waldman et al. (2014) identified aging mechanisms for Lithium ion batteries. Most of these factors are internal to the battery, such as the chemical reaction kinetics, corrosion and loss of water. However measuring these factors in daily life is cumbersome and thus the failure rate of the batteries in a market cannot be predicted accurately in a practical manner.

To create a reliable model for forecasting, Geurts et al. (1996) highlighted that the quality of data is of paramount importance. There has been a lot of literature about the using the POS information to forecast the demand, but as Keifer (2010) points out, forecasting based on POS suffers from a retrospective analysis bias. Furthermore, during new product introductions this approach is not applicable as there is no historical data. Additionally, Keifer (2010) introduces new approaches to forecast new product introductions and web based services. However, there is no discussion about using a multivariate approach or identifying correlations between multiple physical variables and demand for physical products such as replacement batteries.

Multiple studies have been conducted to determine the age of batteries. These studies can be divided into three major categories:

- Experimental studies
- Computational studies
- Combination of experimental and computational studies
These categories were determined based on the tools used for these studies as they impact the results. In subsequent sections in this literature review, these three approaches will be discussed. Additionally, the approaches to handle data to forecast are reviewed and discussed.

2.1 Models to predict the age of lead-acid battery

*Various Mechanisms of Aging*

*Ruetschi (2004)* provides a summary of the aging mechanism and the impact of various mechanisms on battery-life. Additionally, the significance of each aging mechanism and the impact of each mechanism on the various types of lead-acid batteries is determined:

- **Anodic corrosion**: This is the natural aging mechanism of positive plates. This mechanism is mostly common in automotive batteries and stand-by batteries. Additionally, this mechanism is accelerated by battery misuse.

- **Positive mass degradation**: Batteries subjected to cycling such as city buses which make frequent stops and short trips can cause a shallow discharge cycle and thus degrade the positive mass. The positive mass will become softer and will shed.

- **Irreversible formation of lead sulfate**: This mechanism can occur in batteries subjected to higher temperature and/or in the batteries which have a slow discharge rate for a lengthy duration.

- **Short Circuit**: This mechanism is common in automotive batteries and in train-lighting batteries where the usage conditions can be harsh.

- **Loss of water**: This mechanism is common in batteries exposed to higher temperature.
Although Ruetschi (2004) studies several aging mechanisms in detail, a quantitative understanding of the impact of temperature or temperature exposure is not established.

Experimental Studies

In this section the experimental studies from three different researchers are discussed in detail. These researchers have compared and predicted battery life and studied aging behavior in Li-ion and lead acid batteries.

Doerffel and Sharkh (2006) performed experimental studies to predict the remaining battery life. They also compared the results from experimental studies to the existing standard of determining battery capacity empirically by Peukert’s equation, which relates the battery capacity to the discharge current. Based on the result, it was determined that Peukert’s equation is applicable only for constant battery discharges and if the battery discharge rate is variable the Peukert’s equation underestimates the capacity.

In a research article by Thomas et al. (2014), the effects of temperatures on the aging behavior of cycled lithium-ion batteries are investigated quantitatively by electrochemical methods and post-mortem analysis. The results are that temperature dependent aging mechanisms are found by Arrhenius plots, that the different aging mechanism are proven by post-mortem analysis and that the reason for the different mechanisms is found by testing with reference electrodes. All of these results combined confirm that temperature plays an integral role in batteries life cycle (Kouba, 2014). One limitation to Thomas et al. study is that it is focused on Lithium-ion batteries, so it is difficult to apply the results to all automotive batteries. Another limitation is that the
sample consisted of a small number of batteries. The correlations may have been different if batteries had different conditions at a time when testing.

Lu et al. (2014) identified the factors influencing the life cycle of lead-acid batteries in small electric vehicles. The result was that the battery performance and the cycle life improved when the following four methods were used: the combination of grid alloys, mixing paste and curing process parameters control, the selection of the negative organic additives and the sets mode of the positive and negative plates. These results explicitly show that there are many variables to consider when predicting the life cycle of batteries.

**Computational Models**

Computational models are needed as battery aging is irregular and complicated, thus the aging mechanism cannot be replicated. In this section the heuristic model to determine the battery life and some improvements to the basic heuristic model are discussed.

Schiffer et al. (2006) argued that determining the lifetime of a lead-acid battery is complicated because of the irregular operating conditions and the complexity in replicating those conditions experimentally. Hence, a heuristic model is developed, taking into account the impact of various aging mechanisms. Additionally, the results of the model were verified against existing results to validate the model. This model can be used as a systems model for various battery type and operating conditions. Various input parameters of the model include, battery temperature (which is assumed to be ambient temperature for specific conditions), aging mechanism (such as corrosion model and degradation) and state of charge current. Based on the results and by
comparing them with existing data, this model can be used to determine the lifetime of different battery types; however it can be further refined for conditions where the operating current is higher than 10 Ah (ampere-hour).

Esfahanian, Torabi and Mosahebi (2008) refined the model by using computational fluid dynamics (CFD) and Equivalent Circuit Model (ECM) techniques. This model is better due to the fast computation time and greater accuracy from previous models.

Combination modelling approach

In addition to the heuristic model, the modelling approaches can simulate physicochemical mechanisms and consider the incremental decrease in life due to each aging mechanism to predict the battery life. These approaches are discussed in this section.

Sauer and Wenzl (2008) further studied different modelling approaches and provide pros and cons of various approaches. Three different modelling approaches are created:

- Physicochemical aging model: This model includes the aging mechanism of the battery to simulate the battery life. Each mechanism is simulated and the battery life is predicted. This modeling approach is the most complex due to the immense input conditions needed. However, the benefit of this approach is that this could be translated very easily across the various battery types.

- Weighed Ah aging model: This is a heuristic model based on the systems design as performed by Schiffer et al. (2006). This model does not provide any avenues for
continuous improvement to battery manufacturers. However, this is a very powerful model in terms of speed of results.

- Event-oriented aging model: This model is based on the understanding of incremental loss due to each failure mode. This is challenging as the expectation of this model is to quantify each failure mode.

2.2 Connecting Point of Sale (POS) to Forecasting

In the previous section various approaches to determine the battery life were discussed. However, these approaches are not based on any easily measurable physical characteristics and are difficult to determine. Hence forecasting approaches are needed to determine the battery life. In this section various forecasting approaches being employed to determine the demand are discussed.

*Michael et al. (1996)* answered five specific questions for guiding any study. First, who collected the data? Second, why were the data collected? Third, are the sales time series reasonable, consistent, and logical? Fourth, how were the data gathered? Fifth, are the sales figures based on a sample or census? The important issue for forecasters is to know the limitations of the data and any biases that might exist in the data. They suggested there are a few distortions in company generated data due to company politics such as sales quota, tax handling, accounting method, etc. As a result, adjustments to data are required to improve sales forecast. This research concludes that we have to consider the quality of the data used to forecast as well as the models used to make forecasts.
Keifer (2010) identified the weakness in using POS data for forecasting due to the historical nature of these forecast model. Another weakness identified is that they do not work for new products. Demand signals, pre-order sites, prediction markets, gift registries, wish lists, search engines and Web-site usage analysis are suggested as methods to determine the demand of new products. However, these methodologies are applicable to internet based products and/or for new products and are not transferable to products sold in brick and mortar stores.

William et al. (2014) determined that using POS data improves the forecast accuracy. In their study they evaluated the demand of a consumable product. The orders from retailer to suppliers and retailer’s POS data was analyzed and they concluded that the POS data is more related to actual demand of consumers than retailer orders, showing retailer’s orders weren’t actual responses to the market demand. The forecasting with POS data was shown to outperform other approaches by up to 125%. One critical gap of this approach is that POS data don’t include too much information other than the number of unit sold. However, POS data could be very useful if they are incorporated with other important variables.

2.3 Conclusion: The need for a multivariate model of POS data

Based on the literature review, several models exist to predict the lifetime of an individual battery. They can be broadly classified into heuristic, physicochemical and event-based. However, they are difficult to apply to an entire market of batteries in real life as some of the input parameters (such as corrosion, water loss or short-circuit) are difficult to measure on a daily
basis. Additionally, there is no clear connection of these parameters with the external environment such as temperature which is easier to measure and monitor.

Also, although there are several approaches to predict the demand for web based services and for new products, there is little information on the approaches for predicting sales of products sold in brick and mortar stores. Thus, there is a clear need to create a multivariate model to understand the relationship with external conditions such as temperature and battery life.
3. Methodology, Data and Analysis

3.1 Overall Methodology

To determine the impact of temperature on the sales, we followed the three steps illustrated in Figure 3-1. The first step, Data Collection, entails identifying the appropriate level of detail for sales data i.e. whether we should consider sell-in or sell-out data (defined in the table 3-1 below). In addition, this step also involves gathering temperature information. The second step, Data Analysis, involves visualizing the sales data and identifying the most important Stock Keeping Unit (SKU) (sub-group) for further analysis. Finally, in the third step, Data Modeling, the SKU identified in the second step is studied with the temperature information collected in the first step. Thus, the impact of temperature on sales can be quantitatively studied.

**Figure 3-1. Process Flow**

3.2 Data collection

Many companies use a variety of sales data to forecast their sales. As more supplier chains are connected, there are several sales processes even within one chain. Sales data can be divided into two major categories depending on the type of sales information: Sell-in and Sell-out. Sell-
in data represents sales orders from a manufacturer to a retailer. Sell-out data represents sales orders from a retailer to an end customer. Both data are meaningful to understand the current business status and set up the future strategies. Table 3-1 summarizes the benefits of both approaches.

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Sell-in</th>
<th>Sell-out</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retailer → Manufacturer</td>
<td>End customer → Manufacturer</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Benefits</th>
<th>Sell-in</th>
<th>Sell-out</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify volume of the first article production</td>
<td>Able to see the response of end consumers</td>
<td></td>
</tr>
</tbody>
</table>

Because the aim of this research is to improve the sales forecast accuracy, it is more closely related to the behavior of end consumers. The best way to understand the behavior of end consumers is POS data analysis. POS data is considered the most useful Sell-out data.

**Sales information**

POS information captures the sales information on the retailer and customer end. Many companies use POS to manage sales, optimize inventory, maintain customer relationships and etc. Most importantly, POS allows us to understand sales patterns and popular items in different regions and time by real time data. However, sometimes such data are diverse and can have
multiple dimensions such as locations, SKU’s and retailer relationships. These dimensions make it difficult to identify patterns appropriately unless specific data has been identified. SKU’s are based on their usage in particular automobiles and thus can have different sales patterns based either on geography or on climatic conditions. Thus, a particular SKU needs to be identified in order to understand the relationship between sales and temperature without confounding other variables and patterns of SKU’s.

Various dimensions of point of sales (POS) data:

Current POS data includes different components such as vendors, date of sales, zip code, SKU and units sold as indicated in Figure 3-2.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>JCI Fiscal Week Date</td>
<td>Segment Description</td>
<td>Group</td>
<td>Zip Code</td>
<td>Gross Unit Sales</td>
</tr>
<tr>
<td>1048559</td>
<td>6/22/2014 00:00 PASSENGER LIGHT TRUCK/SUV</td>
<td>34</td>
<td>33312</td>
<td>7</td>
</tr>
<tr>
<td>1048560</td>
<td>6/22/2014 00:00 PASSENGER LIGHT TRUCK/SUV</td>
<td>H7</td>
<td>8020</td>
<td>2</td>
</tr>
<tr>
<td>1048561</td>
<td>6/22/2014 00:00 PASSENGER LIGHT TRUCK/SUV</td>
<td>34</td>
<td>46901</td>
<td>3</td>
</tr>
<tr>
<td>1048562</td>
<td>6/22/2014 00:00 PASSENGER LIGHT TRUCK/SUV</td>
<td>121R</td>
<td>32746</td>
<td>3</td>
</tr>
<tr>
<td>1048563</td>
<td>6/22/2014 00:00 PASSENGER LIGHT TRUCK/SUV</td>
<td>75</td>
<td>31516</td>
<td>2</td>
</tr>
<tr>
<td>1048564</td>
<td>6/22/2014 00:00 PASSENGER LIGHT TRUCK/SUV</td>
<td>75</td>
<td>47006</td>
<td>2</td>
</tr>
<tr>
<td>1048565</td>
<td>6/22/2014 00:00 PASSENGER LIGHT TRUCK/SUV</td>
<td>75</td>
<td>35404</td>
<td>1</td>
</tr>
<tr>
<td>1048566</td>
<td>6/22/2014 00:00 PASSENGER LIGHT TRUCK/SUV</td>
<td>65</td>
<td>31008</td>
<td>4</td>
</tr>
<tr>
<td>1048567</td>
<td>6/22/2014 00:00 PASSENGER LIGHT TRUCK/SUV</td>
<td>65</td>
<td>43952</td>
<td>2</td>
</tr>
<tr>
<td>1048568</td>
<td>6/22/2014 00:00 PASSENGER LIGHT TRUCK/SUV</td>
<td>75</td>
<td>45885</td>
<td>2</td>
</tr>
<tr>
<td>1048569</td>
<td>6/22/2014 00:00 PASSENGER LIGHT TRUCK/SUV</td>
<td>75</td>
<td>37924</td>
<td>5</td>
</tr>
</tbody>
</table>

Figure 3-2. POS data composition
The various components illustrated in Figure 3-2 provide information about various parameters. For example, date of sales shows consumers’ buying patterns on a temporal basis. Zip code shows different buying patterns geographically.

Figure 3-3 and Figure 3-4, show the sales of various SKU’s over time and geography. From this our intention is to select one SKU with high sales as well as geographical prevalence.
Figure 3-3. SKU sales by time
Figure 3-4. Zip code of sales

In this thesis POS data from three different vendors is included, thus increasing the data set and also covering the entire geographical US. However, analyzing all the combined data is not only cumbersome but also less impactful, as various SKU may behave differently. Additionally, some SKU may not be geographically prevalent and thus information from these SKU may not be applicable to understand the impact of temperature. Thus, the most meaningful SKU needs to be identified to perform further analysis.

**SKU identification: top 10 sales, widespread location**

To identify the relevant SKU, it is desirable to select the most useful and representative data among all the SKUs. Our main criterion in selecting the SKU was that it should have large enough
dataset to ensure that the model was reliable. Another criterion was to ensure that the selected SKU was widespread enough geographically in order to incorporate temperature diversity and thus understand the impact of temperature. Thus, our rationale was that geographic spread would indicate temperature diversity and create a robust model. This dataset includes more than 30 different group sizes and we assumed that very few people buy more than 2 types of batteries (SKU's). It is assumed that each household will use batteries of same type as it the population of households having a car and a relatively larger vehicle such a bus would be lower. Therefore, considering the SKU with highest sales, size is indicative of the largest SKU and Step one of data-analysis. With this background, the table below shows the top 10 sales of SKU's. Based on this information, we selected SKU 65 as illustrated in Figure 3-5, for further analysis and to identify geographical prevalence.

Next, in order to visualize the geographical prevalence of SKU 65 we used a visualization software Tableau. Tableau is a visualization and business intelligence software developed Tableau Software Company. Tableau enables visualization of huge data sets and meaningful insights can be derived from this analysis.
We visualized sales information for all the SKU’s across all retailers. This enabled us to identify the SKU with the highest sales and helped us determine whether that particular SKU was prevalent across the US.

As illustrated in Figure 3-6 the sales of SKU 65 are shown geographically and based on the Figure 3-6 we can conclude that SKU 65 is sold throughout the US.

![Geographical sales of SKU 65](image)

**Figure 3-6. Geographical sales of SKU 65**

Finally, the POS data of the SKU 65 for a particular region was aggregated. For example the POS data for Boston region consisted of several zip codes shown in the graph below. This information was aggregated as the climatic conditions in a particular metropolitan area were similar.
City selection: 5 cities based on sales and temperature profile

Based on the empirical information on temperature, 5 metropolitan areas were selected. The selection of the cities was done based on the following criteria:

- Mix of cities with and without temperature variation
- Cities where batteries from the SKU 65 are sold

The following cities were selected as shown in Figure 3-8.

- Los Angeles,
- Boston
- Washington D.C.
- Chicago
- Houston
Normalizing sales

Finally, the aggregated sales information from each metropolitan area needs to be normalized as the sales are dependent on the total vehicles in operation (VIO) in a particular metropolitan area. The fraction of VIO in the specific metropolitan area is determined by the following equation.

\[
\text{Normalized sales} = \frac{\text{Total VIO in USA}}{\text{Total Drivers is USA}} \times \frac{\text{Total Drivers in USA}}{\text{Total US population}} \times \frac{\text{Population of specific Metro Area}}{\text{Unit Sales in specific Metro Area}}
\]
Table 3-2. Normalized sales

<table>
<thead>
<tr>
<th>City</th>
<th>Population(M)</th>
<th>Normalized factor</th>
<th>Total Units</th>
<th>Total Units (Normalized)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston</td>
<td>4.5</td>
<td>3.57</td>
<td>8,073</td>
<td>2,264</td>
</tr>
<tr>
<td>Chicago</td>
<td>9.52</td>
<td>7.54</td>
<td>34,211</td>
<td>4,535</td>
</tr>
<tr>
<td>DC</td>
<td>5.86</td>
<td>4.64</td>
<td>20,813</td>
<td>4,482</td>
</tr>
<tr>
<td>Houston</td>
<td>6.18</td>
<td>4.90</td>
<td>68,855</td>
<td>14,060</td>
</tr>
<tr>
<td>LA</td>
<td>18.2</td>
<td>14.42</td>
<td>89,224</td>
<td>6,187</td>
</tr>
<tr>
<td>Total</td>
<td>44.26</td>
<td></td>
<td>221,176</td>
<td>31,528</td>
</tr>
</tbody>
</table>

Temperature

Temperature data from NOAA for last 5 years (2010 – 2014)

Temperature data for each of the following cities was obtained from 2010 to 2014. The temperature information consisted of maximum and minimum temperature data, since battery failure occurs at temperature extremes. Two levels of aggregation needed to be performed for the temperature data. The first was aggregation from daily to weekly temperatures to correlate with the weekly POS, as our thesis company provided weekly POS data. The second was aggregation across the weather stations in a metropolitan region, as the temperature information consisted of temperatures across these various weather stations. For example, the temperature data for the Boston region consisted of daily temperatures at Foxboro, Logan Airport, and other 23 stations.

Additionally, in order to aggregate, the weekly temperature patterns of these regions were evaluated and it was determined that the temperature patterns of these regions were similar, as illustrated in Figure 3-9. Hence, the average weekly maximum and minimum temperature of these regions was aggregated for the entire Boston area, as shown in Figure 3-10.
Figure 3-9. Temperature profiles of 25 stations in Boston area

Figure 3-10. Average weekly temperature of the entire Boston area

This procedure was performed for all the other metropolitan areas identified.
For the Los Angeles region data from the Mount Wilson, Chilao, Mill Creek and Clear creek stations, shown in Figure 3-11 were removed from the calculation of average temperature. The temperature patterns in these regions were different from other regions as shown in Figure 3-12. These regions include national forest and thus do not have a representative number of automobiles which may require battery sales. Hence, this temperature information can be safely removed without impacting the aggregated sales for this metropolitan area. Figure 3-13 shows the average aggregated maximum and minimum temperature for the LA area.
Figure 3-12. 5 Stations not applicable to temperature aggregation

Figure 3-13. Average weekly temperature of the entire LA area
For Houston, DC and Chicago, temperature showed very similar patterns across the entire stations and didn’t show any anomalies as shown in Figures 3-14, 3-16, 3-18. The average maximum and minimum temperatures are shown in Figures 3-15, 3-17 and 3-19.

Figure 3-14. Temperature profiles of 10 stations in Houston area

Figure 3-15. Average weekly temperature of the entire Houston area
Figure 3-16. Temperature profiles of 15 stations in DC area

Figure 3-17. Average weekly temperature of the entire DC area
Figure 3-18. Temperature profiles of 10 stations in Chicago area

Figure 3-19. Average weekly temperature of the entire Chicago area
3.3 Modeling (data from 2010 -2014)

**JMP**

JMP is a statistical software by SAS and it enables identification of quantitative relationships between variables. This is needed for our research as our aim was to identify the relationship between the sales and temperature.

**Regression Analysis**

The “Fit Model” function of JMP was used to create a regression model. We used the following parameters in the model:

1. **Dependent Variables (Y-parameter):** Normalized sales as a continuous parameter.

2. **Independent Variables (X-parameters):**
   - Minimum Temperature, as a continuous parameter
   - Maximum Temperature, as a continuous parameter
   - Year, as ordinal
   - Quarter, as ordinal

The model was created iteratively by plugging in a combination of X-variables and then checking $R^2$. Secondly, the adjusted $R^2$ was also checked to ensure that the model had an appropriate number of variables. The residuals (Actual-Predicted) by Row were plotted to ensure that there are no patterns. Finally, we used a significance value of 0.05 and based on the p-value in the parameter estimates, all the parameters with p-value >0.05, are removed. The parameters are
removed from the model starting with the higher order parameters (e.g. second order parameters are removed before first order parameters) and then those with the highest p-value.

Various combinations of independent variables were used in the model are shown below in Model 1 through 6.

Model 1:

Predictor Parameters included: TMIN, Quarter, Year, Interaction of TMIN and Quarter, quadratic and cubic effect of TMIN. Also data from first quarter of all years was removed to check if there was any variation due to first quarter sales.

Predicted Parameters: Normalized sales

<table>
<thead>
<tr>
<th>Summary of Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
</tr>
<tr>
<td>RSquare Adj</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>Mean of Response</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Quarter[2Q]</td>
</tr>
<tr>
<td>Quarter[3Q]</td>
</tr>
<tr>
<td>Year</td>
</tr>
<tr>
<td>Average of TMIN</td>
</tr>
<tr>
<td>(Average of TMIN-11.8011)*Quarter[2Q]</td>
</tr>
<tr>
<td>(Average of TMIN-11.8011)*Quarter[3Q]</td>
</tr>
<tr>
<td>(Average of TMIN-11.8011)*(Average of TMIN-11.8011)</td>
</tr>
<tr>
<td>(Average of TMIN-11.8011)<em>(Average of TMIN-11.8011)</em>(Average of TMIN-11.8011)</td>
</tr>
</tbody>
</table>

Figure 3-20. Diagnostics of Model 1
Model 2:

Parameters included: TMIN, Quarter, Year, Interaction of TMIN and Quarter, quadratic and cubic effect of TMIN

Predicted Parameters: Normalized sales

| Term                                  | Estimate  | Std Error | t Ratio | Prob>|t| |
|---------------------------------------|-----------|-----------|---------|------|
| Intercept                             | -7422.553 | 1458.841  | -5.09   | <.0001*|
| Quarter[1Q]                           | 13.580965 | 2.256572  | 6.02    | <.0001*|
| Quarter[2Q]                           | -7.279265 | 1.775078  | -4.10   | <.0001*|
| Quarter[3Q]                           | -9.702471 | 3.31191   | -2.93   | 0.0036*|
| Year                                  | 3.6911282 | 0.724865  | 5.09    | <.0001*|
| Tmin                                  | 1.0795884 | 0.219283  | 4.92    | <.0001*|
| (Tmin-7.84418)*Quarter[1Q]           | 1.9073062 | 0.343222  | 5.56    | <.0001*|
| (Tmin-7.84418)*Quarter[2Q]           | -0.822085 | 0.235004  | -3.50   | 0.0005*|
| (Tmin-7.84418)*Quarter[3Q]           | -1.972197 | 0.444074  | -4.44   | <.0001*|
| (Tmin-7.84418)*(Tmin-7.84418)        | 0.2077185 | 0.018722  | 11.09   | <.0001*|
| (Tmin-7.84418)*(Tmin-7.84418)*(Tmin-7.84418) | 0.0037157 | 0.000758  | 4.90    | <.0001*|

Figure 3-21. Diagnostics of Model 2

Model 3:

Predictor Parameters included: TMIN, Tmax, Quarter, Year, Interaction of TMIN and Quarter, and quadratic effect of TMIN

Predicted Parameters: Normalized sales
Summary of Fit

| Parameter | Estimate | Std Error | t Ratio | Prob>|t| |
|---|---|---|---|---|
| Intercept | -7278.905 | 1509.524 | -4.82 | <.0001* |
| Quarter[1Q] | 16.941799 | 2.210039 | 7.67 | <.0001* |
| Quarter[2Q] | -5.665866 | 1.796066 | -3.15 | 0.0017* |
| Quarter[3Q] | -16.4209 | 3.09545 | -5.30 | <.0001* |
| Year | 3.6191208 | 0.750281 | 4.82 | <.0001* |
| Tmax | -0.450572 | 0.432796 | -1.04 | 0.2985 |
| Tmin | 2.3367081 | 0.443009 | 5.27 | <.0001* |
| Quarter[1Q]*(Tmin-7.84418) | 1.6647715 | 0.351325 | 4.74 | <.0001* |
| Quarter[2Q]*(Tmin-7.84418) | -0.960866 | 0.242112 | -3.97 | <.0001* |
| Quarter[3Q]*(Tmin-7.84418) | -1.039331 | 0.411295 | -2.51 | 0.0125* |
| (Tmin-7.84418)*(Tmin-7.84418) | 0.1584872 | 0.016245 | 9.76 | <.0001* |

Model 4:

Predictor Parameters included: TMIN, Tmax, Quarter, Year, Interaction of TMIN and Quarter, and quadratic effect of TMIN.

Predicted Parameters: the square of normalized sales was predicted, instead of just the normalized sales.

Summary of Fit

| Parameter | Estimate | Std Error | t Ratio | Prob>|t| |
|---|---|---|---|---|
| Intercept | 0.505329 | | | |
| RSquare Adj | 0.491996 | | | |
| Root Mean Square Error | 15.10133 | | | |
| Mean of Response | 23.20997 | | | |
| Observations (or Sum Wgts) | 382 | | | |
### Parameter Estimates

| Term                      | Estimate | Std Error | t Ratio | Prob>|t| |
|---------------------------|----------|-----------|---------|-------|
| Intercept                 | -491795.6| 161333.9  | -3.05   | 0.0025* |
| Quarter[1Q]               | 947.31326| 203.2903  | 4.66    | <.0001* |
| Quarter[2Q]               | -258.2701| 149.8081  | -1.72   | 0.0855  |
| Quarter[3Q]               | -1531.704| 198.3015  | -7.72   | <.0001* |
| Year                      | 243.91716| 80.19875  | 3.04    | 0.0025* |
| Tmax                      | -34.96315| 45.37297  | -0.77   | 0.4414  |
| Tmin                      | 202.7532 | 46.62874  | 4.35    | <.0001* |
| (Tmin-7.84418)*(Tmin-7.84418) | 9.2462785| 0.941371  | 9.82    | <.0001* |

**Figure 3-23. Diagnostics of Model 4**

**Model 5:**

Predictor Parameters included: TMIN, Quarter, and quadratic effect of TMIN

Predicted Parameters: Normalized sales

### Summary of Fit

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rsquare</td>
<td>0.409177</td>
</tr>
<tr>
<td>Rsquare Adj</td>
<td>0.406234</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
<td>17.37928</td>
</tr>
<tr>
<td>Mean of Response</td>
<td>31.21604</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
<td>1010</td>
</tr>
</tbody>
</table>

**Figure 3-24. Diagnostics of Model 5**
Model 6:

Predictor Parameters included: TMIN, Quarter

Predicted Parameters: Normalized sales

### Summary of Fit

| Term               | Estimate | Std Error | t Ratio | Prob>|t| |
|--------------------|----------|-----------|---------|------|
| Intercept          | 13.399226| 1.08389   | 12.36   | <.0001* |
| Quarter[1Q]        | 15.59257 | 1.324172  | 11.78   | <.0001* |
| Quarter[2Q]        | -9.134888| 1.065725  | -8.57   | <.0001* |
| Quarter[3Q]        | -14.26195| 1.317281  | -10.83  | <.0001* |
| Tmin               | 1.8862826| 0.09515   | 19.82   | <.0001* |

Figure 3-25. Diagnostics of Model 6

Model Discussion

Let us discuss each of the five models we have created.

Model 1: This model shown in the Figure 3-20 provides the best $R^2$ value but as we notice we have used a cubic function of minimum temperature. Additionally the data from 1st quarter for all years was removed to check whether quarter would impact the sales. Based on these results we observe that quarter does impact the results, but removal of sales data from 1st quarter and including a cubic expression of Tmin may not be warranted.
Model 2: In this model shown in the Figure 3-21 the sales data from 1st quarter of all years is included, but the model expression is kept the same from model 1. As the $R^2$ is lower from model 1, this implies that the variation from 1st quarter induces more variability in the overall data. Additionally, we still have a cubic expression and an interaction of temperature and quarter, both of which may be unwarranted.

Model 3: In this model shown in the Figure 3-22 the cubic expression from model 2 is dropped, but the interaction is still kept in the model. Additionally, the expression includes the maximum temperature. Although, the $R^2$ is decent, but the inclusion of interaction and the maximum temperature may not be warranted.

Model 4: In this model shown in the Figure 3-23 the model expression for independent variables is kept the same but the dependent variable, Normalized sales is transformed (squared) to check if it yields a better fit. The fit does not improve, in fact the $R^2$ is reduced and hence the transformation of normalized sales is not justifiable.

Model 5: This model shown in the Figure 3-24 includes the quadratic effect of minimum temperature and does not include either the interaction or the cubic effect. We observe that this model achieves a more discreet $R^2$ value but has the advantage of being more parsimonious, e.g. using less variables. The use of quadratic variable for temperature may be warranted by the fact that the relationship between temperature and the battery life (and sales) is not linear.

Model 6: This model shown in the Figure 3-25 is a further simplification of model 5 and does not include the quadratic effect. We observe that the $R^2$ is further reduced. Also as discussed in model
5, the relationship between battery life(sales) and temperature may not be linear and is proved by the poor fit of this model.

From these 6 models, model 1 and 2 include the cubic and quadratic effect of Tmin as well as the interaction of Tmin and quarter. Hence even though the $R^2$ for these models is higher than 50%, we did not select these models as they may be over fitting due to inclusion of additional variables. Model 3 includes the interaction and model 4 further complicates by transforming the sales. Thus we do not select these models as well. Model 6, on the other hand oversimplifies and only uses the linear relationship between temperature and sales and thus has a lower $R^2$ and predictive power.

From model 5 we can see that the temperature is indeed a predictor of sales. Notice it is the minimum and not the maximum temperature that is the best predictor. A model that uses the minimum temperature, both linear and squared along with the quarter, like model 5 above seems to offer a good compromise between predictive power and parsimony. Hence model 5, was selected from the above 6 models.
4. Validation of the approach

Is there a way to validate the approach used to generate model 5 as a predictor of sales based on temperature? There is one way: to use it with new data. The sponsor company can apply it with new data. This however takes time. Is there a way to validate the approach of model 5 now? There may be a way: to use only part of the data, instead of all the data to generate a model which will then be applied to predict the values in the rest of the data. Data can be segregated for this exercise either in time or in geography.

Thus, two models were created: one with three cities namely Chicago, LA and Houston with data from 2011 to 2014 and another with data from 2011, 2013 and 2014 with all the cities and indicated in Table 4-1. This was done as we wanted not only to create the model but also to validate the model. One option was to use all five cities to develop the model, but then we would have no way to validate it unless we obtained additional data. Instead, we decided to use the data from three cities to create a model, and then use the data from the other two models to validate the model. Additionally, to validate the model across time we decided to use data from three years and then use the data from one year to validate the model.

Table 4-1. Validation of models

<table>
<thead>
<tr>
<th>Model</th>
<th>City</th>
<th>Year</th>
<th>Purpose</th>
</tr>
</thead>
</table>
We chose Chicago, LA and Houston because these cities encompassed the range of minimum and maximum temperatures seen across the five cities as shown in Table 4-2, and thus the model could be used to predict the sales in Boston and Washington D.C.

**Table 4-2: Range of minimum Temperatures across the cities**

<table>
<thead>
<tr>
<th>City</th>
<th>Higher end of minimum Temperature (°C)</th>
<th>Lower end of minimum Temperature (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston</td>
<td>21.7</td>
<td>-14.0</td>
</tr>
<tr>
<td>LA</td>
<td>21.5</td>
<td>4.8</td>
</tr>
<tr>
<td>DC</td>
<td>24.2</td>
<td>-10.5</td>
</tr>
<tr>
<td>Houston</td>
<td>26.1</td>
<td>-1.3</td>
</tr>
<tr>
<td>Chicago</td>
<td>24.0</td>
<td>-17.2</td>
</tr>
</tbody>
</table>

Similarly, the years 2011, 2013 and 2014 were chosen for model B as these years encompassed the range of minimum and maximum temperatures across the four years, as shown in Table 4-3. Hence the data from 2011, 2013 and 2014 could be used to predict the sales in 2011.

**Table 4-3: Minimum and Maximum Temperatures across the years**

<table>
<thead>
<tr>
<th>Year</th>
<th>Maximum Temperature (°C)</th>
<th>Minimum Temperature (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>26.1</td>
<td>-14.0</td>
</tr>
<tr>
<td>2012</td>
<td>25.0</td>
<td>-10.1</td>
</tr>
<tr>
<td>2013</td>
<td>24.9</td>
<td>-14.3</td>
</tr>
<tr>
<td>2014</td>
<td>24.6</td>
<td>-17.2</td>
</tr>
</tbody>
</table>
The first model (Model G) provides an understanding of fit in terms of geography as this was model was developed with data for Houston, Los Angeles and Chicago. The second model (Model T) provides an understanding in terms of time and this model was developed with data from 2011, 2013 and 2014.

4.1 Model G diagnostics

The model diagnostics for Model G are shown Figure 4-1.

Regression Plot

![Regression Plot](image)

Summary of Fit

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
<td>0.405144</td>
</tr>
<tr>
<td>RSquare Adj</td>
<td>0.400187</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
<td>18.45579</td>
</tr>
<tr>
<td>Mean of Response</td>
<td>40.89451</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
<td>606</td>
</tr>
</tbody>
</table>

**Figure: 4-1 Model diagnostics for Model G: R^2**
As illustrated in Figure 4-1, the $R^2$ and adjusted $R^2$ for model G are 40%. This implies that with the variables in the model explain, 40% of the variability in the sales is explained by this model.

The Pareto chart, in Figure 4-2 illustrates the relative significance of each parameter in the model. Figure 4-3 shows that the minimum temperature (Tmin) and the quadratic effect of Tmin are the most important variables in the model.

### Pareto Plot of Transformed Estimates

<table>
<thead>
<tr>
<th>Term</th>
<th>Orthog Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average of TMIN</td>
<td>11.61524</td>
</tr>
<tr>
<td>(Average of TMIN-1.4528)*(Average of TMIN-11.4528)</td>
<td>9.26380</td>
</tr>
<tr>
<td>Quarter[3Q]</td>
<td>1.93990</td>
</tr>
<tr>
<td>Quarter[2Q]</td>
<td>-1.93182</td>
</tr>
<tr>
<td>Quarter[1Q]</td>
<td>-1.20946</td>
</tr>
</tbody>
</table>

![Figure 4-2: Pareto Plot for Model G](image)

Additionally, from Figure 4-3, describing the parameter estimates, it can be observed that all the parameters are statistically significant.

### Parameter Estimates

| Term                                      | Estimate   | Std Error | t Ratio | Prob>|t| |
|-------------------------------------------|------------|-----------|---------|------|
| Intercept                                 | 2.5165218  | 2.100989  | 1.20    | 0.2315 |
| Quarter[1Q]                               | 14.401557  | 1.575935  | 9.14    | <.0001* |
| Quarter[2Q]                               | -8.036456  | 1.324293  | -6.07   | <.0001* |
| Quarter[3Q]                               | -18.52613  | 1.690849  | -10.96  | <.0001* |
| Average of TMIN                           | 2.7718893  | 0.140517  | 19.73   | <.0001* |
| (Average of TMIN-11.4528)*(Average of TMIN-11.4528) | 0.0952339 | 0.007707 | 12.36   | <.0001* |

![Figure 4-3: Parameter Estimates for Model G](image)

Finally the expression in Figure 4-4 describes the quantitative relationship between sales, temperature and quarter. The significance of quarter implies that even though sales are impacted...
by temperature, the impact is also dependent on the quarter. Therefore, for the same minimum
temperature, the sales could vary by quarter. This may indicate that the customer behavior may
be different in quarters or that the mechanism of failure, i.e. physicochemical mechanisms, could
be different in quarters. This further indicates that other climatic factors, such as humidity etc.
or age of the battery, may additionally influence the failure rate. Additionally, another inference
is that the third quarter would have the lowest sales and the first and fourth quarters would have
the maximum sales. However, this could just be a manifestation of the temperature as the low
minimum temperatures during 1st and 4th quarter may trigger the higher sales.

Furthermore, the quadratic effect implies that sales bottom out at a certain temperature and
sales increase at the other temperature extreme. However, as quarter is also a factor in the
model, the temperature at which sales bottom out will different for each quarter.

$$2.51652175403159$$

$$+ \text{Match[Quarter]} \begin{cases} "1Q" \Rightarrow 14.4015568297034 \\ "2Q" \Rightarrow -8.0364561076666 \\ "3Q" \Rightarrow -18.526133413611 \\ "4Q" \Rightarrow 12.1610326915738 \\ \text{else} \Rightarrow . \end{cases}$$

$$+ 2.77188931917744 \times \text{Average of TMIN}$$

$$\left\{ \text{Average of TMIN} - 11.452794878231 \right\} +$$

$$\times \left\{ \text{Average of TMIN} - 11.452794878231 \right\} \times 0.09523385467006$$

Figure 4-4: Prediction Expression for Model G
4.2 Model T diagnostics

As discussed previously comparing the $R^2$ and adjusted $R^2$ provides a measure of the explanation of variability and also the measure of whether the model is overfitted, as shown in Figure 4-5.

![Regression Plot](Image)

**Summary of Fit**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
<td>0.39803</td>
</tr>
<tr>
<td>RSquare Adj</td>
<td>0.393957</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
<td>17.79322</td>
</tr>
<tr>
<td>Mean of Response</td>
<td>31.22096</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
<td>745</td>
</tr>
</tbody>
</table>

*Figure: 4-5 Model diagnostics for Model T: $R^2$*

The prediction equation and the pareto chart are shown below in Figure 4-6 and 4-7.
Based on this it can be concluded that both minimum temperature and the quadratic effect of minimum temperature are more important than quarter. Additionally, we can derive very similar conclusions are from Model T.

**Prediction Expression**

\[
1.52016479823857 + \begin{cases} 
1Q \Rightarrow 15.3359960316323 \\
2Q \Rightarrow -7.5992214579832 \\
3Q \Rightarrow -20.200397225495 \\
4Q \Rightarrow 12.4636226518457 \\
\text{else} \Rightarrow .
\end{cases} + 2.40942411949481 \times \text{Average of TMIN}
\]

\[
\left[ \begin{array}{c}
\text{Average of TMIN - 9.41303712902415} \\
+ 0.08600801465012 \end{array} \right]
\]

**Figure 4-7: Prediction Expression for Model T**
4.3 Insights on the validity of the approach

Based on Figure 4-8 to 4-10 the approach of forecasting sales based on temperature model is more robust across time than across geography. We also notice that for both geography and time, the trends for actual and predicted sales are the same. However, for Model G, the difference between actual and predicted is much higher both for Boston and Washington D.C, than for Model T, the difference between actual and predicted values of 2012. This information can be used to prioritize what data to use when refining the model further. For example, if there is an equal amount of data available for geography or for time then the data from additional geographical locations should be used.

![Figure 4-8: Model validation for Boston (Model G)](image-url)
Figure 4-9: Model validation for Washington D.C. (Model G)

Figure 4-10: Model validation for Year 2012 (Model T)
Additionally, it also implies that there are more variations across geographies and thus a model generated based on data from one region can’t be extrapolated to another region. In this case data from the West coast, South and Midwest regions was used for the model and validated against East coast cities, Boston and Washington D.C. Based on the actual vs. predicted plots this implies that model needs to be built based on region to increase predictability.

Based on the Figures 4-8 and 4-9, we see that the pattern of predicted and actual sales in both Boston and Washington D.C. are similar, but the absolute values are different. Hence we normalized both the predicted and actual sales based on the overall average of predicted and actual sales for the duration. This helped us understand the prediction of change over time.

Based on Figures 4-11 and 4-12, we observe that the %change from average is similar for predicted and actual. This illustrates the fact that the direction and magnitude of change can be predicted by the model, but the absolute value of sales cannot be determined by the model.

This illustrates the fact that the even though the sales were normalized based on the vehicles in operations, but there is still a difference due to factors such as demographics, public transportation and other local preferences.
Figure 4-11: Model validation for Boston (Model G) based on change

Figure 4-12: Model validation for Washington D.C. (Model G) based on change
5. Conclusion and Future Work

In this study, we established a correlation between sales and temperature to explain the variability in battery sales. Based on the results from the model, we found that there is a linear and quadratic relationship between the minimum temperature and battery sales. Additionally, based on the model validation for geography and time we determined that the model is more robust across time than across geography. Thus, this helps prioritize the resources when refining the model by adding additional data.

What this means for our sponsor company is that they will be able to use temperature data to improve their sales forecast. This can be done by developing models that use historical data of the minimum temperature in a region and the point of sales of a given SKU in that region to predict future sales of that SKU as a function of future minimum temperature. The model can be developed using multiple regression, with the quarter, and minimum temperature as predictors. The minimum temperature in the model is related to the sales both linearly and quadratically.

Based on these results, our thesis sponsor can further refine the model by adding sales and temperature information from various geographies. Additionally, another factor such as age of the battery can also be added to further refine the model. The age of the battery can be calculated based on results from a small customer survey in a representative metropolitan area.

This additional understanding of the impact of temperature on the sales forecast allows firms not only to respond quickly to customer needs but also to reduce inventory costs, ultimately increasing their profits. Furthermore, this understanding and improvement in battery failure and thus sales represents a causal factor analysis in improving sales forecasts of automotive batteries.
References


Kevin Kouba (2014). Can climate contribute to battery life expectancy?. *Audiology Online, 1-1*


