Using Historical Data to Predict Future Service

Georgina Fetterhoff

Ferris State University

December 10, 2018

Table of Contents

Page

Chapter 1: Introduction to the Problem…….……3

Chapter 2: Software and Ligature Reviews……...9

Chapter 3: Methodology…….11

Chapter 4: Findings…….22

Chapter 5: Recommendations and Conclusion……23

References……....25

Using Historical Data to Predict Future Service

Chapter 1

Introduction to the Problem

I am researching the significance of historical data when trying to project future work. I feel finding a way to track and predict service of products installed would be benefitable to many different industries that supplies and service products. This could include heating and cooling companies, industrial machinery providers, car dealerships and the industry that I will be looking at, which is the garage door and operator industry. The idea behind this project is to help a company predict service based off sales to help fulfill manpower needs, make advertising more affective and evaluate the performance of the models sold.

Problem Statement

The company from which I am retrieving the data has no way to predict the quantity of service calls they will receive or where in the greater West Michigan area they will be located within a given period of time.

Purpose of the Study

Analysis of historical data can help predict volume and placement of service calls which in turn can help define labor resource need. This predictive data can also give a better idea of where to advertise their services, help identify trends in the models sold, and help differentiate what the most profitable ones are and if there are any defect trends. All are important aspects to help the company adapt and grow with the current economy. The method that will be applied could be used in other companies within the garage door industry or in other similar industries.

Research Question

The questions I am researching are:

- How can analysis of historical data predict future service call volume and location?
    - How can historical data evaluate the performance of models sold?

Resources

The resources needed for this project is Microsoft SQL databases from the Overhead Door Company of Grand Rapids, licensed copy of Microsoft SQL Server 2016, licensed copy of Microsoft Office 365 and a licensed copy of Microsoft Power BI, all being supplied by Overhead Door Company of Grand Rapids for this project. Also needed is a computer to run the database server and Power BI software that will be utilized for the project and for the presentation, being supplied by myself.

Definition of Terms

- Address – Reference to the physical address in which the garage door operator was installed.
- Model Number – The identifying number of a product given to it by the manufacturer.
- Operator – A motorized device that is used to control the up and down motion of a garage door.
- Powerhead – The electric motor of a residential operator.
- Rail – The track that the door is attached to and follows when the motor runs in the operator.
- Unit – Refers to an installed operator and rail.

- Warranty – A guarantee given to the original purchaser that the product being purchased will repaired or replaced if necessary for a defined period as long as there is no sign of negligence or damage by the purchaser or by act-of-god.

Assumptions and Limitations

The assumptions of this study include:

- Service addresses within the database are correct.
- Operator model installed is correct in database.
- All service done on operator was performed by Overhead Door Co. of Grand Rapids.
- Service records are correct and complete.
- Outliers will be correctly identified.
- Database information being pulled from is without errors within its tables.

The limitations of this study include:

- Only one company's data is being used.
- Only covers the greater West Michigan area of the state of Michigan.
- The model of operators that are being researched are only Overhead Door brand.
- Only two models are being researched: Odyssey 1000 Model 7030 and Standard Drive Model 1026.
- Only one industry is being looked at.
- Software limitations include the use of Microsoft SQL Server 2016, Microsoft PowerBI, Microsoft Excel and Sage 300 ERP because this is the software the company has licensing for and where their data is utilized and maintained.

- Historical data is limited in the time frame of January 1, 2007 to December 31, 2017 and current data to use to compare results is limited to January 1, 2018 to August 31, 2018.
- Outliers will include incomplete addresses and addresses outside a 50-mile radius of the city of Grand Rapids Michigan.

Project Plan

Table 1

Project Plan

Due	**Item**
9/14/18	Project Proposal
9/14/18	Obtain copy of databases
9/21/18	Create new database and import scrubbed data.
9/28/18	Research regression techniques.
9/28/18	Status Report 1
10/5/18	Decide on regression technique to be used.
10/12/18	Find outliers and remove from database.
10/19/18	Complete regression and finalize data for projection.
10/26/18	Status Report 2
10/26/18	Connect database to Power BI
11/2/18	Basic dashboard created.
11/9/18	Drill down and main functionality running in dashboard.

11/16/18	Status Report 3
11/16/18	GIS Map created in dashboard.
11/23/18	Work on presentation of dashboard. Finalize paper.
12/3/18	Final Project Due w/Presentation

Risk Management

There are several risks that could keep this project from being completed. Below is a chart of risks and possible solutions.

Table 2

Risk Management

Risk	**Contingency**
Illness	• Take off from work to allow more time to complete project. • Try to keep ahead of schedule to allow extra time.
Problems with projection (how/math/kind)	• Seek help from Dr. Gogolin or previous professors. • Seek out colleagues or other students that might help. • Internet resources such as YouTube, onlinestatbook.com, coursera.org or other colleges.
Work related issues	• Speak with owner about rearranging schedule. • Speak with Dr. Gogolin about work issues and time.
Database problems	• Internet resources such as YouTube, codementor.io, w3schools.com, or technet.microsoft.com • Speak with Dr. Gogolin or Professor Emerick.
Computer problems	• Use laptop as a backup

Time management	• Keep a schedule of activities to help stay on task. • Try to work ahead so there is extra time for unseen issues.
Data problems	• Utilize USPS.com to get correct address information. • Manual fixes for invalid or incomplete data. • Manual inspection of data.

Mock-Up of Presentation

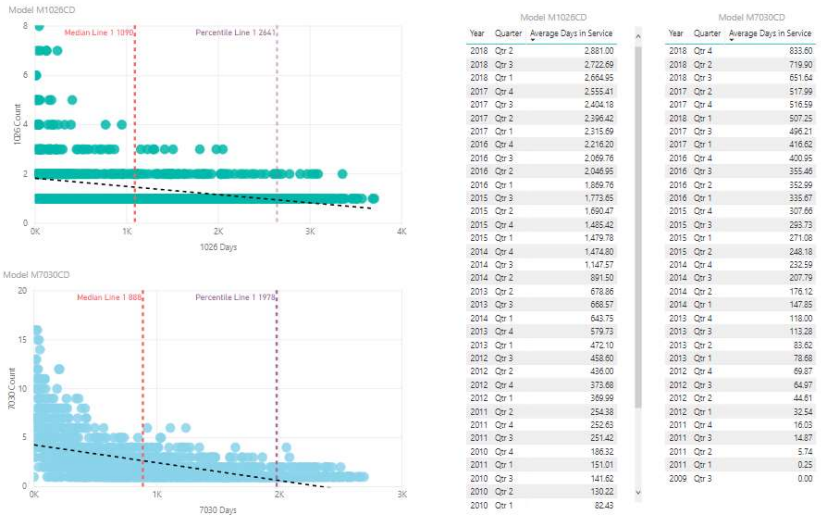


Figure 1. PowerBI Dashboard. (Microsoft, 2018)

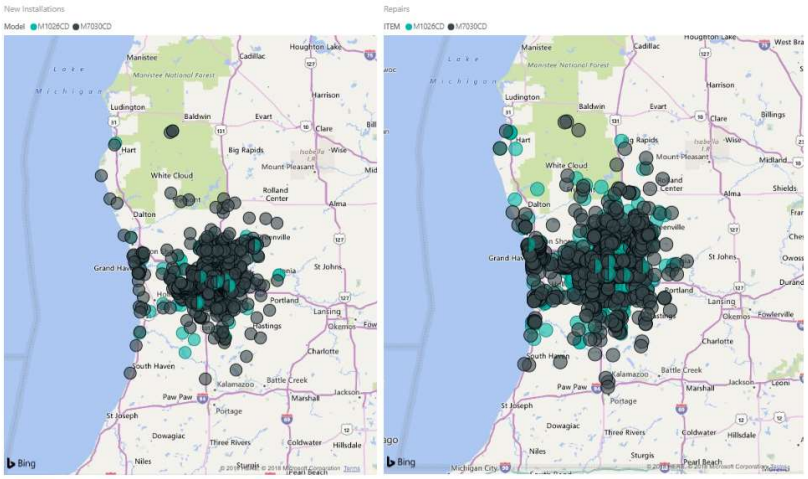


Figure 2. PowerBI Dashboard. (Microsoft, 2018)

Presentation Evaluation Criteria

Success of the project will be determined by having a hundred percent working dashboard and the prediction of service has been created or if prediction is not possible an explanation of why and a plan on what should be done differently to make it possible. The accuracy of the prediction hopefully will be within the eighty-percentile range. Also, the paper written, and actual presentation should be accounted for within the evaluation. I will be presenting my presentation online with a narrative of how the project was completed and a walkthrough of the working dashboard.

Chapter 2

Software Reviews

Microsoft Power BI. Power BI is available as software-as-a-service, desktop application and as a traditional report server for on-premise (Microsoft, 2018). The cloud option is what I concentrated on because this is the version of the software Overhead Door Co. of Grand Rapids has licensing to. The software connects easily to the Microsoft SQL database and Microsoft Excel for seamless integration. Power BI offers an interactive, real-time dashboard that is easy to share on PC, tablets and phones via apps (Microsoft, 2018). Data can be transformed within Power BI and has advanced analytics with the familiarity of Excel (Microsoft, 2018), making learning and adapting easier. Power BI also allows you to create maps using Bing maps or importing maps from ArcGIS. The ease of integration with Office 365 makes it easy for the company to use and understand.

Tableau. Tableau is available as software-as-a-service, desktop application, and as a traditional server application (Tableau, 2018). The desktop version is what I concentrated on based off the availability to obtain a student version of the software. Tableau offers the ability to

connect directly to a Microsoft SQL database and offers interactive dashboards (Tableau, 2018). Like Power BI, Tableau is able to process higher amounts of data and will connect to hundreds of different types of data sources (Tableau, 2018).

Literature Reviews

Local linear regression for estimating time series data. This paper is about utilizing local linear regression (LLR) and is a method that is quite a mathematical technique. It takes LLR and compares it to five other models used for time series (Nottingham & Cook, 2001). The take-away of the paper is that LLR is useful when there isn't a lot of history available or where changes and shifts in the processes occur (Nottingham & Cook, 2001). Overall, I feel that this method is not what is needed for the project see that I have multiple years of data available.

Geographic information systems: A mixed methods spatial approach in business and management research and beyond. This journal article discusses using geographic information systems (GIS) in research to help give the data spatial context (Frels, Frels, & Onwuegbuzie, 2011). It is stated that interactive GIS applications can add value to researchers and the mix quantitative and qualitative data can aid in participant enrichment (Frels et al., 2011). Using GIS in research that has geographical features allows the researcher to see the data in a different way. I feel it is an easy visual way to present data to an audience that may not have data experience or even well-developed technical skills. Everyone is exposed to maps in school and in everyday life so presenting information in a familiar and visually appealing way can be very useful in this project.

Linear regression models for prediction of annual heating and cooling demand in representative Australian residential dwellings. This journal article though not exactly on the same subject line as my project does offer a good look at using linear regression. I also like the

layout of the paper and format. The emphasis of the paper is on multiple regression coefficients (Aghdaei, Kokogiannakis, Daly, & McCarthy, 2017). This type of regression is useful if you are looking into the relationship between multiple independent variables and a single dependent variable. This type of regression is not needed given the data used for this project contains only one dependent and one independent variable.

Ordinary Least Squares Regression Method Approach for Site Selection of Automated Teller Machines. This paper uses ordinary least squares regression to predict optimum locations for future automated teller machines (ATMs) (Bilginol, Denli, & Zafer Eker, 2015). The model used is most likely what I will be using in this project. They show how least squares help minimize error and use the ArcGIS Spatial Statistics toolbox to achieve the results (Bilginol et al., 2015). Utilizing the data they had and the ArcGIS tool they calculated that the locations decided where within five percent error (Bilginol et al., 2015). This method is intriguing, but I do not have a license or access to the ArcGIS system and the company in which this project is being done for does not currently have nor will have licensing for this software system.

Chapter 3

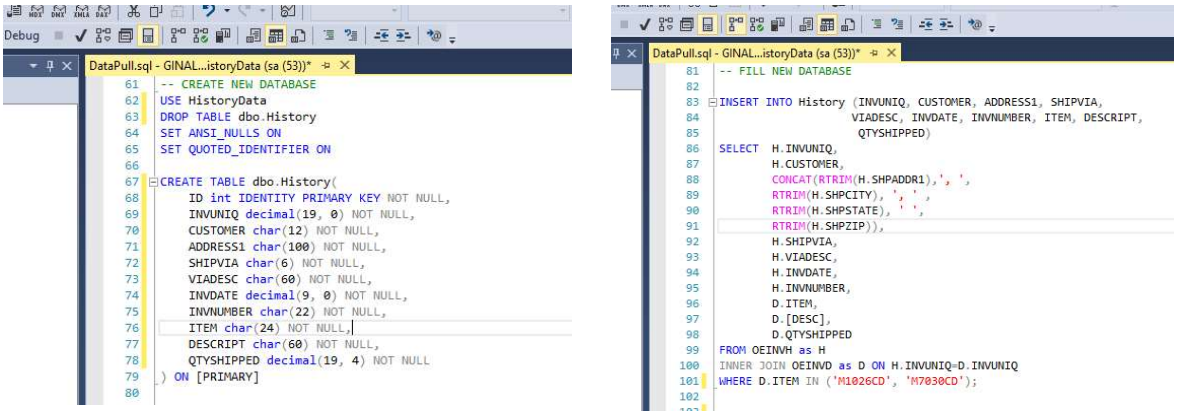
Methodology

First a test sample was taken from the data source to inspect. Upon inspection the tables and columns in the Sage 300 ERP the data will be taken from the OEINVH table which contains the order entry invoice header and the OEINVD table which contains the order entry invoice detail information. The columns that will be exported out are as follows and can be seen on the right side of Figure 3:

- INVUNIQ – The primary key for the record.
- CUSTOMER – The account number for the record.

- SHPADDR1 – The ship to address for the record.
- SHPCITY – The ship to city for the record.
- SHPSTATE – The ship to state for the record.
- SHPZIP – The ship to zip code for the record.
- SHIPVIA – The ship via code that identifies the type of record. For this project we are looking at SH0000 new residential installations, SH00003 residential repairs and SH0007 no charge warranty work.
- VIADESC – The name associated with the SHPVIA.
- INVDATE – The date that the record was invoiced, also used as the installation date.
- INVNUMBER – The invoice number created for the sale.
- ITEM – The part number for an inventory item sold. For this project we are looking at two models of residential garage door operators: M1026CD and M7030CD.
- DESC – Description of the ITEM.
- QTYSHIPPED – The quantity on the ITEM sold to the CUSTOMER.

Then a warehouse database structure was drawn out and created in SQL to house the information exported. Note in Figure 3 on the right side that the data is being combined for the address taking the database out of first normal form. This is to make sorting the records by address easier and for use in plotting the addresses into a GIS map. Figure 3 left side shows SQL statement used for the creation of the new database and then the right side shows the SQL statement to fill the new database with the sample data.



```
-- CREATE NEW DATABASE
USE HistoryData
DROP TABLE dbo.History
SET ANSI_NULLS ON
SET QUOTED_IDENTIFIER ON

CREATE TABLE dbo.History(
    ID int IDENTITY PRIMARY KEY NOT NULL,
    INVUNIQ decimal(19, 0) NOT NULL,
    CUSTOMER char(12) NOT NULL,
    ADDRESS1 char(100) NOT NULL,
    SHIPVIA char(6) NOT NULL,
    VIADESC char(60) NOT NULL,
    INVDATE decimal(9, 0) NOT NULL,
    INVNUMBER char(22) NOT NULL,
    ITEM char(24) NOT NULL,
    DESCRIPT char(60) NOT NULL,
    QTYSHIPPED decimal(19, 4) NOT NULL
) ON [PRIMARY]
```

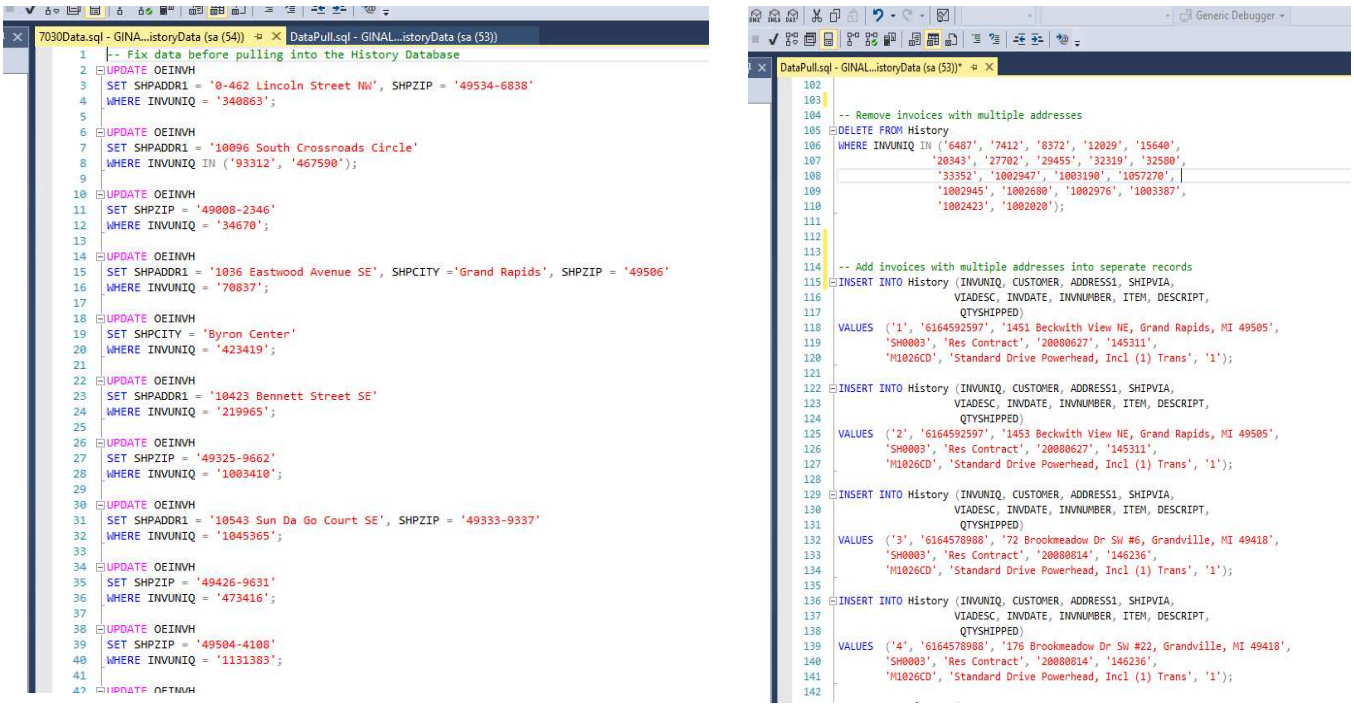
```
-- FILL NEW DATABASE

INSERT INTO History (INVUNIQ, CUSTOMER, ADDRESS1, SHIPVIA,
                     VIADESC, INVDATE, INVNUMBER, ITEM, DESCRIPT,
                     QTYSHIPPED)
SELECT  H.INVUNIQ,
        H.CUSTOMER,
        CONCAT(RTRIM(H.SHPADDR1),', ',
        RTRIM(H.SHPCITY), ', ' ,
        RTRIM(H.SHPSTATE), ' ',
        RTRIM(H.SHPZIP)),
        H.SHIPVIA,
        H.VIADESC,
        H.INVDATE,
        H.INVNUMBER,
        D.ITEM,
        D.[DESC],
        D.QTYSHIPPED
FROM OEINVH as H
INNER JOIN OEINVD as D ON H.INVUNIQ=D.INVUNIQ
WHERE D.ITEM IN ('M1026CD', 'M7030CD');
```

Figure 3. SQL Statements for creating & filling database ("Microsoft SQL Server," 2016)

Once the database was filled, the data was then exported to Microsoft Excel in order to look for errors and outliers. It was at this time the realization of how poorly the data had been entered in the main source. Errors included improper, incomplete, multiple and misspelled addresses along with using the improper SHIPVIA code for the type of work was also found. Figure 4 shows an example of a street that was typed into the system in four different ways. The correct name of the street is South Crossroads Circle SE, having it entered multiple ways makes comparing new installs to service work done extremely hard. With the data in Excel the errors were noted and then SQL statements were made to correct the errors to make comparing the address field simple and error free, Figure 5. Outliers that were not within a 50-mile radius of Grand Rapids Michigan were also removed.

48	72117	6163648121	10085 S. Crossroads Circle, Caledonia, MI 49316	SH0003	Res Contract	201
49	72570	6163648121	10087 S. Crossroads Circle, Caledonia, MI 49316	SH0003	Res Contract	201
50	65213	6163648121	10094 Crossroads Circle SE, Caledonia, MI 49316	SH0003	Res Contract	201
51	467590	6162751100	10096 Crossroads Circle SE, Caledonia, MI 49316	SH0003	Res Contract	201
52	93312	6163648121	10096 Crossroads, Caledonia, MI 49316	SH0003	Res Contract	201
53	53943	6163648121	10098 S Crossroads Circle SE, Caledonia, MI 49316	SH0003	Res Contract	201
54	337357	6168783603	101 Brewer Park Circle SE, Byron Center, MI 49315	SH0003	Res Contract	201

Figure 4. Example of a street that has been improperly inputted.("Microsoft SQL Server," 2016)



```
-- Fix data before pulling into the History Database
UPDATE OEINVH
 SET SHPADDR1 = '0-462 Lincoln Street NW', SHPZIP = '49534-6838'
 WHERE INVUNIQ = '340863';

UPDATE OEINVH
 SET SHPADDR1 = '10096 South Crossroads Circle'
 WHERE INVUNIQ IN ('93312', '467590');

UPDATE OEINVH
 SET SHPZIP = '49008-2346'
 WHERE INVUNIQ = '34670';

UPDATE OEINVH
 SET SHPADDR1 = '1036 Eastwood Avenue SE', SHPCITY ='Grand Rapids', SHPZIP = '49506'
 WHERE INVUNIQ = '70837';

UPDATE OEINVH
 SET SHPCITY = 'Byron Center'
 WHERE INVUNIQ = '423419';

UPDATE OEINVH
 SET SHPADDR1 = '10423 Bennett Street SE'
 WHERE INVUNIQ = '219965';

UPDATE OEINVH
 SET SHPZIP = '49325-9662'
 WHERE INVUNIQ = '1003410';

UPDATE OEINVH
 SET SHPADDR1 = '10543 Sun Da Go Court SE', SHPZIP = '49333-9337'
 WHERE INVUNIQ = '1045365';

UPDATE OEINVH
 SET SHPZIP = '49426-9631'
 WHERE INVUNIQ = '473416';

UPDATE OEINVH
 SET SHPZIP = '49504-4108'
 WHERE INVUNIQ = '1131383';
```

```
-- Remove invoices with multiple addresses
DELETE FROM History
 WHERE INVUNIQ IN ('6487', '7412', '8372', '12029', '15640',
                   '20343', '27702', '29455', '32319', '32580',
                   '33352', '1002947', '1003190', '1057270',
                   '1002945', '1002680', '1002976', '1003387',
                   '1002423', '1002020');



 -- Add invoices with multiple addresses into seperate records
INSERT INTO History (INVUNIQ, CUSTOMER, ADDRESS1, SHIPVIA,
                     VIADESC, INVDATE, INVNUMBER, ITEM, DESCRIPT,
                      QTYSHIPPED)
 VALUES  ('1', '6164592597', '1451 Beckwith View NE, Grand Rapids, MI 49505',
         'SH0003', 'Res Contract', '20080627', '145311',
         'M1026CD', 'Standard Drive Powerhead, Incl (1) Trans', '1');

INSERT INTO History (INVUNIQ, CUSTOMER, ADDRESS1, SHIPVIA,
                     VIADESC, INVDATE, INVNUMBER, ITEM, DESCRIPT,
                      QTYSHIPPED)
 VALUES  ('2', '6164592597', '1453 Beckwith View NE, Grand Rapids, MI 49505',
         'SH0003', 'Res Contract', '20080627', '145311',
         'M1026CD', 'Standard Drive Powerhead, Incl (1) Trans', '1');

INSERT INTO History (INVUNIQ, CUSTOMER, ADDRESS1, SHIPVIA,
                     VIADESC, INVDATE, INVNUMBER, ITEM, DESCRIPT,
                      QTYSHIPPED)
 VALUES  ('3', '6164578988', '72 Brookmeadow Dr SW #6, Grandville, MI 49418',
         'SH0003', 'Res Contract', '20080814', '146236',
         'M1026CD', 'Standard Drive Powerhead, Incl (1) Trans', '1');

INSERT INTO History (INVUNIQ, CUSTOMER, ADDRESS1, SHIPVIA,
                     VIADESC, INVDATE, INVNUMBER, ITEM, DESCRIPT,
                      QTYSHIPPED)
 VALUES  ('4', '6164578988', '176 Brookmeadow Dr SW #22, Grandville, MI 49418',
         'SH0003', 'Res Contract', '20080814', '146236',
         'M1026CD', 'Standard Drive Powerhead, Incl (1) Trans', '1');
```

Figure 5. SQL Statement corrections ("Microsoft SQL Server," 2016)

Once that process was completed the data was again exported into Excel, this time into two separate workbooks. One with the SHIPVIA code of SH0000 for new installs and one with SHIPVIA codes SH0003 for repairs and SH0007 for no charges or warranty. The spreadsheet that contained the installs had two new columns added. The first is DAYSNSVC for number of days from the installation date to the service date for each service order completed. The other is an ID column to match the service orders to the installations. Using Excel, the columns in the installation worksheet were matched and filled with the information needed, creating new rows for each service order found. For the installation record the DAYSNSVC column is fill with a zero to show that it is the record of the installation and not a service record.

The next step was to create a new database that would hold the new formed data and then the data was imported into the database from the Excel worksheet using the import feature in Microsoft SQL Server Management Studio 17. Initially this transformation was attempted in SQL Statements but errored out continuously from how the initial data types were setup. The

date filled was setup as plain text and in YYYYMMDD order and SQL did not calculate the number of days, instead we got simple numeric difference. Example of this error is 20080101 subtracted from 20121201 in days is 1,796 but the numeric difference is 41,100. This made for inaccurate data for the linear regression and had to be dealt with in a slightly more time-consuming matter.

The data in the Excel sheet was used to create the DAYSNSVC column first by taking the existing INVDATE column and using the Text to Columns function to create the new INSTDATE column that has the format of MMDDYYYY. Then Excel will calculate the math of subtracting one date from another which fills the DAYSNSVC column with a difference in days and not a general numeric difference. Once all the records had the DAYSNSVC column calculated the data was flattened (plain text answer replacing the formula) the database was removed and the recreated with the new column and the data was once again imported.

The next step is to create scatter plots for each model and evaluate the trendline and intercept. The simplest way to perform this was in Excel seeing that we already have the data in a worksheet. We utilize the Insert Chart function and select scatter plot and the data we want plotted and Excel creates a chart. The initial scatter plots simply looking at the days in service were not linear or helpful, Figure 6.

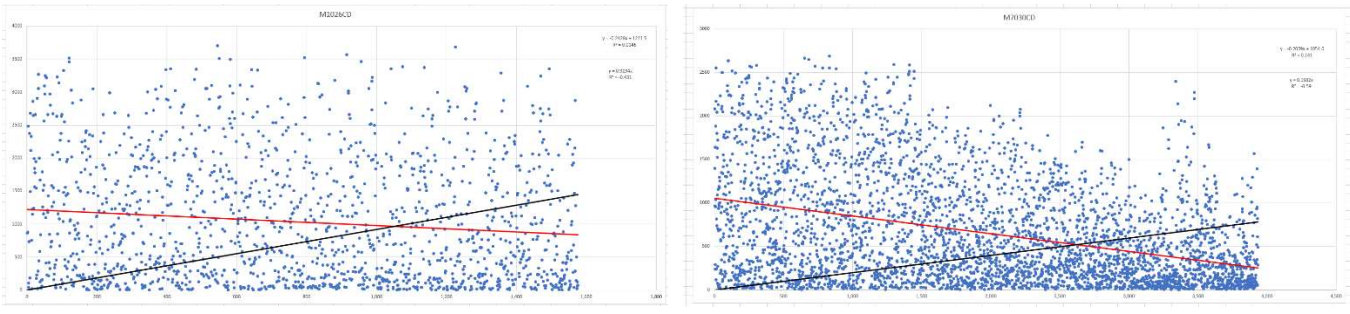


Figure 6. First scatter plots based off only days in service. ("Microsoft Excel," 2016)

Again, data was pulled from the database this time using the COUNT function to count how many units fall under each DAYSNSVC. The retrieved data was then put into Excel and another set of scatter plots were created, Figure 7.

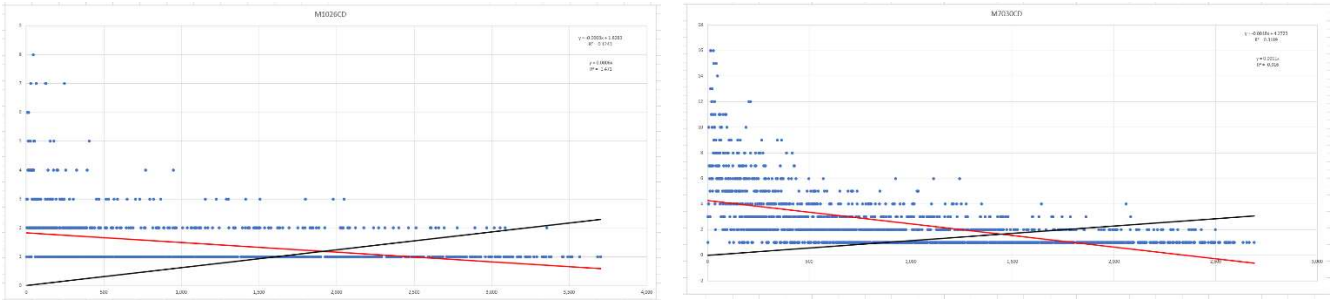


Figure 6. First scatter plots based off only days in service. ("Microsoft Excel," 2016)

Excel will place the trendline and Y-Intercept on the chart along with giving you the linear regression formula, R^2 value, and the Y-Intercept value. These values are useful for comparing to the Data Analysis Tool to make sure you did not make an error in selecting the data from the scatter plot to the Summary Output.

Now we look state a hypothesis to test. This will either pass or fail the data for use in projection. The hypothesis we are going to use will be 'The count of units to be serviced goes up with the number of days in service'. In other words, the older a unit gets the more likely it will need to be serviced. If this holds true than the data will be able to predict a timeframe in which a new installation will need service. $H_0\ \mu =$ Unit counts increase as days in service increases, this is the null hypothesis. The alternate hypothesis is $H_1\ \mu\ \neq$ Unit counts increase as days in service increases.

Next we need to state our alpha level, we are going to use .05 or 5% which is the standard in the scientific community (PennState Eberly College of Science, 2018). Using the Data Analysis: Regression function in Excel we ran the Regression Statistics for each model, M1026CD and M7030CD.

Looking at the Summary Output for the M1026CD model below in Figure 7 we can use the Coefficients Intercept as "a" the y-intercept and X Variable 1 as "b" the slope to find the linear regression equation of $Y = a + bX$ making it $Y = 1.8283 + -0.0003X$ (PennState Eberly College of Science, 2018). Looking at the R^2 value given of 0.1243 we know that a prediction has a 12.43% chance of falling on the prediction line. The P-value shows as 7.8E-252 which is an extremely small number and is much less than our .05 alpha level and therefore we reject the null hypothesis (PennState Eberly College of Science, 2018).

Regression Statistics								
Multiple R	0.352626529							
R Square	0.124345469							
Adjusted R Square	0.123555167							
Standard Error	0.813006795							
Observations	1110							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	103.9980694	103.9980694	157.3392	7.6193E-34			
Residual	1108	732.3658946	0.660980049					
Total	1109	836.363964						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.828332958	0.040695638	44.92700096	7.8E-252	1.748483749	1.908182167	1.748483749	1.908182167
X Variable 1	-0.000332533	2.65104E-05	-12.54349217	7.62E-34	-0.000384549	-0.000280517	-0.000384549	-0.000280517

Figure 7. Regression statistics for model M1026CD. ("Microsoft Excel," 2016)

Looking at Figure 8 which is the Regression Statistics for the model M7030CD we can create the linear regression equation as $Y = 4.2723 + -0.0018X$. The R^2 value shows that 31.89% of predictions would fall on the predication line and the P-Value is 2.3E-134 which is also extremely small making us reject the null hypothesis for this model also. Out of the two models the M7030CD had a better regression line but there still not a close enough relationship between the independent and dependent variable to make the data viable for prediction.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.564724093
R Square	0.318913301
Adjusted R Square	0.318483593
Standard Error	1.787542008
Observations	1587

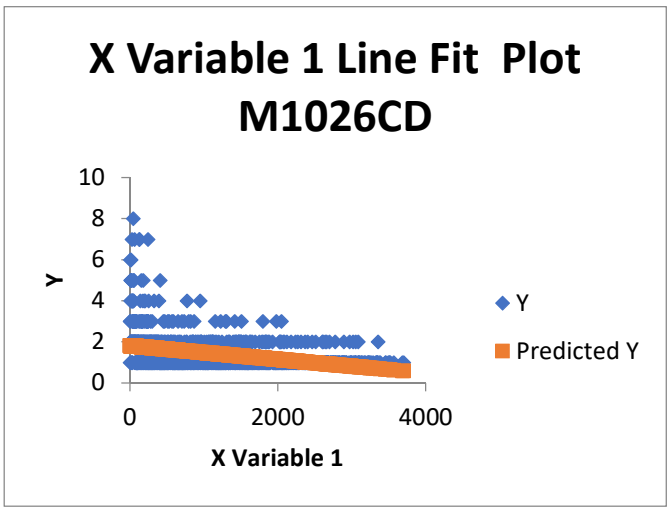
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2371.439306	2371.439306	742.1633	2.3032E-134
Residual	1585	5064.560694	3.195306432		
Total	1586	7436			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	4.272319839	0.079688587	53.61269423	0	4.116013719	4.428625959	4.116013719	4.428625959
X Variable 1	-0.001822819	6.69104E-05	-27.24267505	2.3E-134	-0.001954061	-0.001691576	-0.001954061	-0.001691576

Figure 8. Regression statistics for model M7030CD. ("Microsoft Excel," 2016)

Utilizing the graphs that Excel also gives to us in the Data Analysis, Figure 9, we can also see that for both models the prediction falls into the negative which is not a probability.



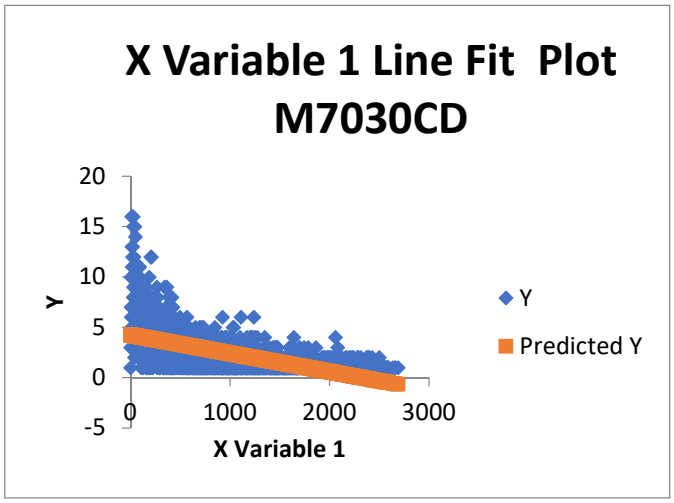


Figure 9. Regression statistics Prediction graphs. ("Microsoft Excel," 2016)

Since prediction is not probable with the data set we have we can only build a dashboard that utilizes what we do know. First we log into PowerBI online through our Office 365 portal. Next you can either use an existing workspace or create a new one. Once you are in a workspace you click on Create+ and create a new dashboard. Before you can begin to make dashboard visualizations you have to either connect to a database or import data into PowerBI. Both the

database and the counts data in Excel are imported into PowerBI for presentation using the Get Data arrow at the bottom left of the page. Once the data has been imported you can make visualizations using that data. The first page we will shows the scatter plots with the trend line, median line and percentile line, Figure 10.

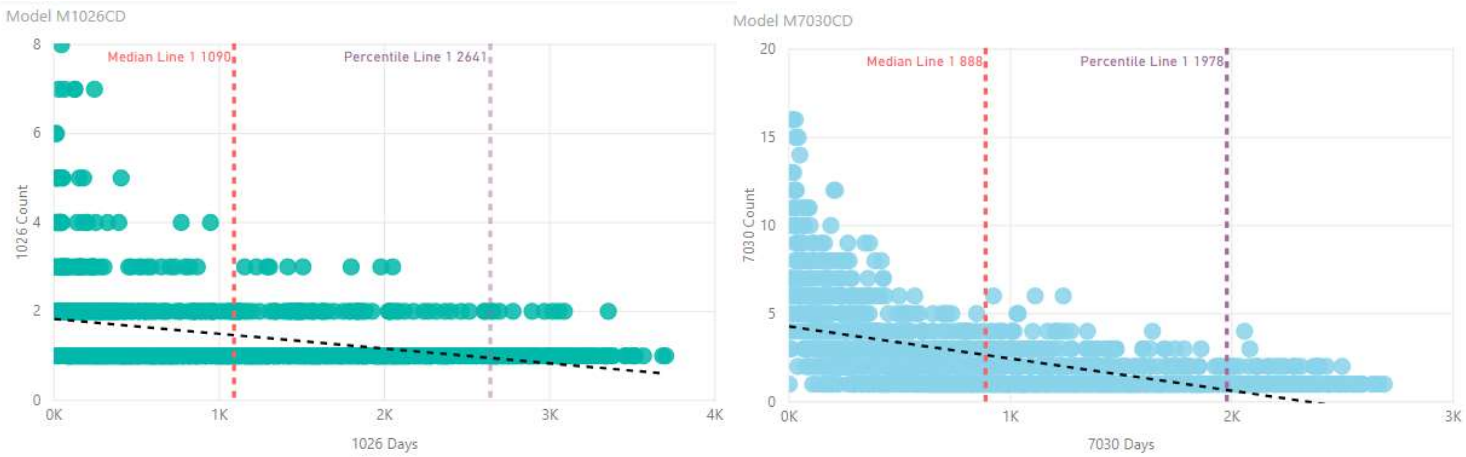


Figure 10. PowreBI scatter plots. (Microsoft, 2018)

Once you click on the Scatter Plot icon you drag and drop from the data list the columns you want to use for the X and Y Axis. We are using the number of days for the X Axis and the count of calls for the Y Axis. Then from the Analytics panel we choose to show the Median, Trend Line and a Percentile Line for 90%. After that the format and colors are set in the Format table and the graphs look like the ones in Figure 10. From these graphs we can see that 90% of the model M1026CD installed fails before reaching 2,641 days or 7.46 years and the model M7030CD 90% fails 1,978 or 5.62 years from installation.

Now we will create tables on the dashboard, Figure 11, that show the average days in service quarterly. Under Visualizations we selected the Table icon, drag and drop from the data source the columns we want to look at which is Average Days in Service and use the Install Date to determine the Quarter and Year. You can title and change the look of the table using the

Format table and that gives us the tables in Figure 11. The final Data page in the PowerBI dashboard now looks like Figure 12.

Model M1026CD

Year	Quarter	Average Days in Service
2010	Qtr 1	82.43
2010	Qtr 2	130.22
2010	Qtr 3	141.62
2010	Qtr 4	186.32
2011	Qtr 1	151.01
2011	Qtr 2	254.38
2011	Qtr 3	251.42
2011	Qtr 4	252.63
2012	Qtr 1	369.99
2012	Qtr 2	436.00
2012	Qtr 3	458.60
2012	Qtr 4	373.68
2013	Qtr 1	472.10
2013	Qtr 2	678.86
2013	Qtr 3	668.57
2013	Qtr 4	579.73
2014	Qtr 1	643.75
2014	Qtr 2	891.50
2014	Qtr 3	1,147.57
2014	Qtr 4	1,474.80
2015	Qtr 1	1,479.78
2015	Qtr 2	1,690.47
2015	Qtr 3	1,773.65
2015	Qtr 4	1,485.42
2016	Qtr 1	1,869.76
2016	Qtr 2	2,046.95
2016	Qtr 3	2,069.76
2016	Qtr 4	2,216.20
2017	Qtr 1	2,315.69
2017	Qtr 2	2,396.42
2017	Qtr 3	2,404.18
2017	Qtr 4	2,555.41
2018	Qtr 1	2,664.95
2018	Qtr 2	2,881.00
2018	Qtr 3	2,722.69

Model M7030CD

Year	Quarter	Average Days in Service
2009	Qtr 3	0.00
2011	Qtr 1	0.25
2011	Qtr 2	5.74
2011	Qtr 3	14.87
2011	Qtr 4	16.03
2012	Qtr 1	32.54
2012	Qtr 2	44.61
2012	Qtr 3	64.97
2012	Qtr 4	69.87
2013	Qtr 1	78.68
2013	Qtr 2	83.62
2013	Qtr 3	113.28
2013	Qtr 4	118.00
2014	Qtr 1	147.85
2014	Qtr 2	176.12
2014	Qtr 3	207.79
2014	Qtr 4	232.59
2015	Qtr 1	271.08
2015	Qtr 2	248.18
2015	Qtr 3	293.73
2015	Qtr 4	307.66
2016	Qtr 1	335.67
2016	Qtr 2	352.99
2016	Qtr 3	355.46
2016	Qtr 4	400.95
2017	Qtr 1	416.62
2017	Qtr 2	517.99
2017	Qtr 3	496.21
2017	Qtr 4	516.59
2018	Qtr 1	507.25
2018	Qtr 2	719.90
2018	Qtr 3	651.64
2018	Qtr 4	833.60

Figure 11. Average days in service tables from PowerBI. (Microsoft, 2018)

Both models had a recall on their boards in the 1st quarter of 2011 and the recall lasted until the 4th quarter or 2013. After the initial recall the model M7030CD continued to fail and had a second recall on the board used in it which ended in the 2nd quarter or 2014. We can see from the tables that the M1026CD has become a much stronger performing unit since the first recall and as of the 3rd quarter of 2018 has over twice the performance lifespan.

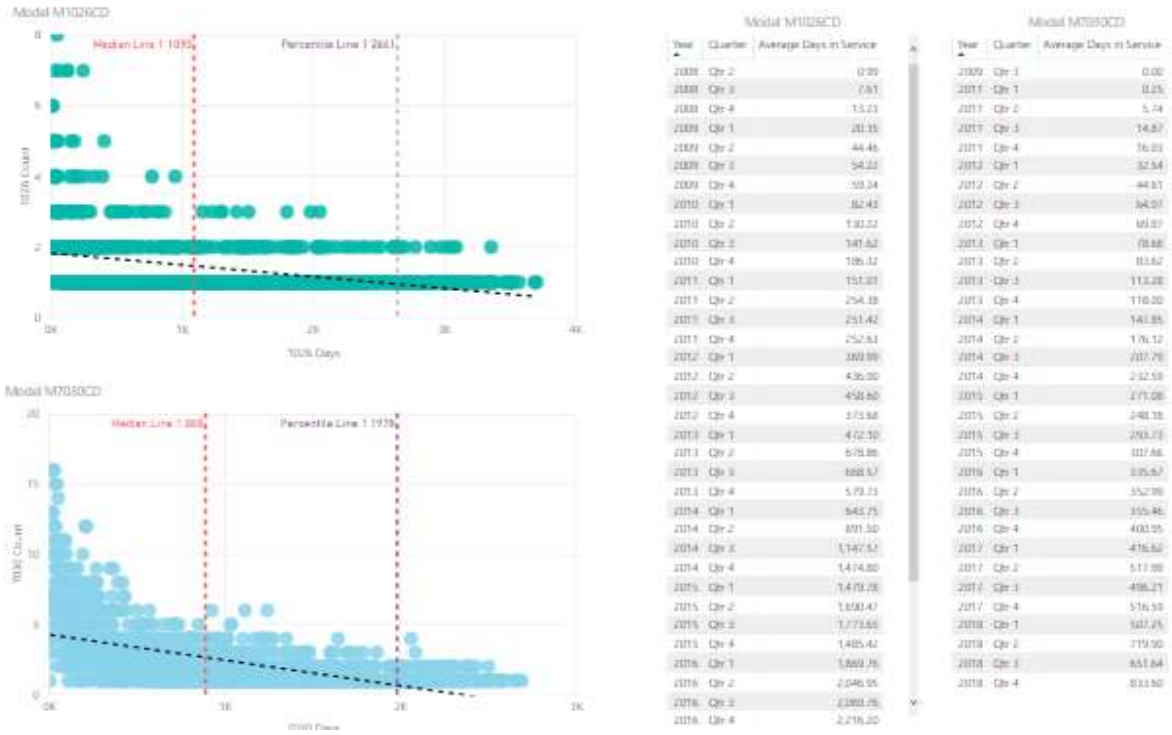


Figure 12. Finished Data page from PowerBI. (Microsoft, 2018)

The second page is created to show maps of the installations and service, Figure 13. At the bottom of the Dashboard there is a yellow plus sign, when clicked it creates a new blank page. On this new page we will select the Map icon and then drag and drop the ADDRESS1 column from the database into the location setting. We set the Legend by dropping ITEM into the column and that then separates the addresses shown on the map into two colors one for each model. You can utilize the Format table to change the colors and title of the map. Also if you notice any addresses that are out of the area that you want to concentrate on you can click on them and choose exclude and the map will remove that record.

You will notice that there are service points showing up in areas that do not show installations. Again, this shows that the data retrieved from the main database has errors and inconstancies that need addressed for a valid predication to be made.

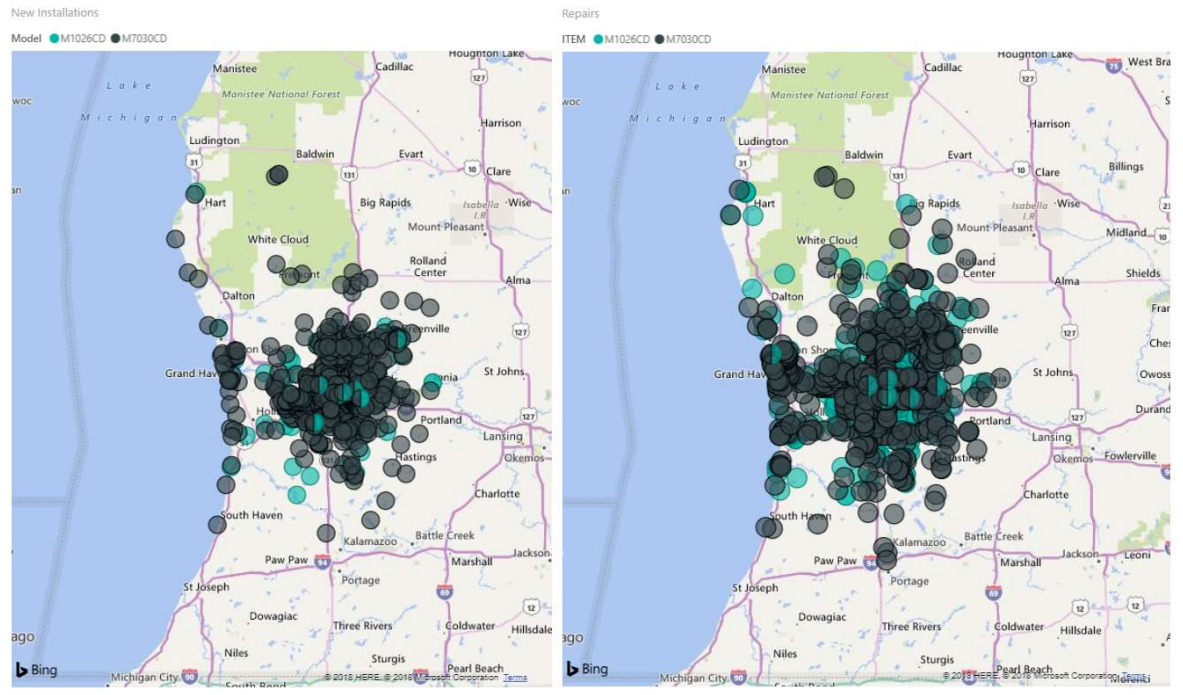


Figure 12. Maps of address for installs and service from PowerBI. (Microsoft, 2018)

Chapter 4

Findings

Utilizing real world data can be challenging. Unlike clean straight forward data that is used in academia there are many factors that makes for a messy dataset. These factors include the hardest component to combat which is the human factor. This project if completed will run far over the given time frame and had to be cut short due to time restrictions. The conclusion on the existing data state is that there are many things that need to be researched and fixed in order to get more realistic results. Looking into some of the service records that show addresses that there is no install record for unveiled even more errors that how an address is entered. Sales have been entered into the wrong sales category making it impossible to find the installs. There are also many entries that had multiple addresses listed or only the numeric portion of the address entered which make correctly identifying them improbable without research. We chose twenty random service calls and pulled the original paperwork for the installation and all service that we

could find. This showed that even the service calls are sometimes wrongly categorized. Sometimes a customer stated the door was not working so the call was made in the Residential Door Service category, but once a tech went out to the job they found that the door works fine it is the operator that is not working. At the time of billing the person doing the billing did not change the category and therefore the service call stays in a state of error. This showed true for categorizing service as Commercial in lieu of Residential, Door in lieu of Operator and vice versa.

It is unfortunate and disappointing that the database has such an array of errors. One positive note from the project is that the company has new rules, checks and balances, and auditing in place to make the data moving forward much more reliable. However, the company did find the information on the Average Day in Service very interesting and useful. The Model 7030 shows as the weaker performer, yet it is the more expensive machine with a longer warranty. Model 1026 is considered a builder grade opener yet outperforms. It is believed that this is due the recall on the boards, from the data shown and service calls that were researched the Model 1026 board issues has been resolved whereas the Model 7030 is still receiving service calls from board issues.

Chapter 5

Recommendations and Future Work

After sitting down with the President of the company and going over the findings of the database errors and how to move forward it has been determined that this project will be picked back up in one year. It was decided to be too costly to continue to try to fix the historical data and it is best to look at the data moving forward. Along with putting in the new techniques of how to enter addresses and how to find proper address, the billing department is now verifying

that the job category is correct before invoicing. On top of this all new accounts are being looked over weekly and fixed from any errors before permanent records are made.

Moving forward it has been suggested to utilized the serialized inventory function that the Sage 300 ERP software has built-in. Utilizing serial numbers would make matching up sales and service much cleaner in not only looking at the history of Operators but also Doors. Another recommendation is to make sure that the operator model is written down on the job order when installation is done. There were some instances where a different model was substituted due to lack of inventory but was not recorded properly when billed.

For anyone wanting to do this type of project it is recommended that you make sure your data is correct and error free. You should allot yourself extra time to clean and investigate a sample section of the data to confirm that you have a good representation of what you are wanting to predict. Also you should always check the P-Value and the R^2 early on to see if the data is a good candidate for prediction.

References

Aghdaei, N., Kokogiannakis, G., Daly, D., & McCarthy, T. (2017). Linear regression models for prediction of annual heating and cooling demand in representative Australian residential dwellings. *Energy Procedia*, *121*, 79–86. https://doi.org/10.1016/j.egypro.2017.07.482

Bilginol, K., Denli, H., & Zafer Eker, D. (2015). Ordinary Least Squares Regression Method Approach for Site Selection of Automated Teller Machines. *Procedia Environmental Sciences*, *26*, 66–69. https://doi.org/10.1016/j.proenv.2015.05.026

Frels, J. G., Frels, R. K., & Onwuegbuzie, A. J. (2011). Geographic information systems: a mixed methods spatial approach in business and management research and beyond. *International Journal of Multiple Research Approaches*, *5*(3), 367+. Retrieved from http://go.galegroup.com.ezproxy.ferris.edu/ps/i.do?ty=as&v=2.1&u=lom_ferrissu&it=DIourl&s=RELEVANCE&p=AONE&qt=SN~1834-0806~~VO~5~~SP~367~~IU~3&lm=DA~120110000&sw=w

Microsoft. (2018). Microsoft Power BI. Retrieved September 14, 2018, from https://powerbi.microsoft.com/en-us/get-started/?&OCID=AID719832_SEM_bHb24t0B&lnkd=Google_PowerBI_Brand&gclid=CjwKCAjwuO3cBRAyEiwAzOxKsj3K9kZglZTI_KKCW9hCMxUsSm1qLnomjWYbETbze78i2HD5KXxyiBoCD8gQAvD_BwE

Microsoft Excel. (2016). Microsoft.

Microsoft SQL Server. (2016). Microsoft.

Nottingham, Q. J., & Cook, D. F. (2001). Local linear regression for estimating time series data. *Computational Statistics & Data Analysis*, *37*(2), 209–217. https://doi.org/10.1016/S0167-9473(01)00006-8

PennState Eberly College of Science. (2018). Lesson 1: Simple Linear Regression | STAT 501. Retrieved November 3, 2018, from https://onlinecourses.science.psu.edu/stat501/node/250/

Tableau. (2018). Tableau Software. Retrieved September 14, 2018, from https://www.tableau.com/