PennState

Data Visualization of Public Construction-Related Data
Dental Dojo
Project Report

December 14, 2020

Submitted to: Dental Dojo
Nate Parkinson
Salt Lake City, Utah

Submitted by: Epic Energy Team
Joseph Sepich
Myung Joon Kim
Yixuan Wang
Manasvi Mittal
Shunqi Zhang
Hanzhong Ye

Advisor: Marc Rigas

EXECUTIVE SUMMARY

Today, energy is the material basis for the survival and development of human society. Energy has a particularly important strategic position in the national economy. Our Sponsor, Epic Energy, is a California-based energy efficiency home improvement organization, who takes great responsibility to educate and improve customers' homes to help people conserve and consume energy more efficiently.

In order to provide a platform for our sponsor to make better strategic business decisions, this project will implement a solution that will allow Epic Energy to harness the mass amounts of publicly available building permit data, which is a format that they are familiar with. In this project, we will make a software tool that will integrate and present information from California building permits and then put this onto a map. This map interface will visualize the types of construction projects performed in the sponsor's target home markets. Our goal is to have this tool enable the sponsor to identify underserved locations in the solar and home energy efficiency space.

We'll complete the entire project according to a sequential timeline. The team will gather raw data, perform data reformation, implement a data storage solution, perform exploratory data analysis (EDA), create end user analytics, rapidly develop a front-end prototype, and produce a website to host the tool. The sponsor will receive tangible deliverables to track our progresses that will include raw permit data, a hosted database, an interim analytics report, a final report, a hosted website, a technical developer guide, a website prototype, and a final video. The major cost of this project will be purchasing cloud services for hosting the data storage system and the website.

TABLE OF CONTENTS

Contents

EXECUTIVE SUMMARY ii
TABLE OF CONTENTS iii
1.0 INTRODUCTION 1
2.0 PROBLEM STATEMENT 1
3.0 PROJECT OBJECTIVES 2
4.0 PROJECT MANAGEMENT / APPROACH 3
4.1 Statement of Work 2
5.0 BENEFITS 11
6.0 DELIVERABLES 12
7.0 TEAM CAPABILITIES 16
8.0 BUDGET NARRATIVE 17
REFERENCES 3
APPENDIX A: PERSONNEL VITAS 4
APPENDIX B: PROJECT BUDGET 18

1.0 INTRODUCTION

Various technological inventions have brought human civilization to an unprecedented height. At the same time, the energy consumed by human beings is also increasing day by day, among which coal and oil are the main energy sources today. Today, energy is the material basis for the survival and development of human society and has a particularly important strategic position in the national economy. Energy is equivalent to the blood of the city, and it drives the operation of the city. The higher the degree of modernization, the stronger the dependence on energy, because energy maintains the following important functions: lighting, transportation, catering, heating, cooling, and automated management systems.

Our sponsor, Epic Energy, takes great insight and responsibility into energy nowadays to help people conserve and consume energy more effectively and efficiently while saving you money and adding value. They are located at western coast mainly California. They have run this business for more than 4 years. Though they are not as large as LG Solar USA, they are good and ambitious people with dreams, down-to-earth and hard work. This can be seen through the sponsor representative Nathan Parkinson (CMO), who merged had his previous firm acquired by Epic Energy, since they were already providing the company with a large majority of their customers. With his background in marketing Nate's first idea for a use case of this tool was as to determine who would be their target customers to market their services to.

While this is one good use case for the building permit data our main goals will be to synthesize the public data and serve it in multiple analytics on a web server. We do not want to narrow our goal to merely a marketing tool, but a tool that the construction firm can use to determine underserved areas, cross reference construction/customers with demographics, and monitory the activities of rival firms. All of these could be accomplished through the proper use of the building permit data.

2.0 PROBLEM STATEMENT

The construction industry is a large, old industry that creates numerous amounts of data, especially public data, via required building permits that are filed with local government. These permits are often required to be made public via some sort of open data laws, which stands true for the state of California. Currently most contractors and construction firms fail to utilize the mass amounts of data and ever-growing technological capabilities. The project will aim to implement a solution that will allow Epic Energy to harness publicly available data to improve their business. Since Epic Energy's main source of business takes place in California, there is an opportunity to take advantage of the state's open data laws with building permits.

A large barrier to entry in terms of using this data is the heterogeneity. Although the state of California requires that building permits be made public, they do not specify the format and do not aggregate this data, so the data resides on each municipality's web sites, which also have different formats. The availability and ease of use of this data from the perspective of inserting data into a relational database varies greatly. We can see very desirable formats from jurisdictions such as Sacramento Valley, which have their data hosted in ArcGIS in a csv or API format (https://tinyurl.com/yxesenpw). Others, such as the city of Sacramento (https://tinyurl.com/y4fv6n9g) have multiple excel documents, which although don't contain all the data in one place, they are easily synthesized. The worst-case scenario, which seems to occur in many of the more rural jurisdictions only list the data as PDFs, which are very difficult to extract the data from. An example of this is the city of Redding (https://tinyurl.com/yxoclu5c). It is quite clear that the synthesis of all these data sources is a must in delivering an actionable product to the sponsor.

Once this challenge of data gathering and wrangling is conquered data analysis can take place. This portion is vital to providing a useful tool for Epic Energy. The company can use this data to understand what projects competitors are completing and focus their advertising efforts on perspective customers.

3.0 PROJECT OBJECTIVES

There are three main project objectives for this project to be considered success for our sponsor. The first is to create a method/process for gathering public building permit data, so the company can make use of this public data. The second objective would be to create an analytical tool to assist the company in determining target customers for installation of their energy solutions. The third objective would be to win an award for the quality of our final presentation of the project. Below are some definitions for the objectives.

A proper data gathering process will include data sources, data wrangling code, and an end data format description. The data sources will include websites based in the counties listed in the deliverables section, the wrangling code will involve a repeatable process for the patterns found in this data, and the end data format description will come in the form of a SQL statement. This will enable further developers to gather more data to expand the tool and this objective would do most of the heavy lifting in terms of data gathering.

The analytical tool developed will be based on further research. When developing this tool our objective is to find some clever or unique approach to take to assist us in our final objective. In the deliverable section an interim report is described that will act as an update to this objective to use preliminary analysis to dictate the end goal.

Combining the data gathering and analysis into a final report should meet the final objective of having a phenomenal final report that shows off the research direction that was taken.

4.0 APPROACH

There are three main task categories: data gathering, data analysis, and user interface development. Data gathering includes tasks such as investigating data sources, downloading data, tidying data, and inserting data into a data storage solution. Data analysis will involve extracting information from the data gathered. This will include exploratory data analysis, exploring potential analytical tools, writing an interim report, and working with interface developers to sync interface functionality with analysis. The interface development portion includes tasks such as developing a basic interface for accessing the data, creating the website to be hosted, creating/presenting the data in a useful visualization, and presenting the sponsor with iterative prototypes.

4.1 Data Synthesis

Data synthesis mainly involved gathering the raw data, which includes both finding the data sources and downloading the data that they contained. The county list proved by the sponsor was used as a priority when deciding which data to gather. Using a compiled list of links, the raw data was extracted based off the format of each data source. Additional to the building permit data sources utility information usage and demographic information was pulled from PG&E and the 2010 Census respectively, since the 2020 Census had not yet been finalized.

The second part to the data synthesis process was converting the raw data into the same format. This task was completed by Oct 23 and began with wrangling the data into a tidy format. A tidy format is an important specification as many PDF reports for building permits do not meet this requirement and must be translated. During transformation of the data the features included in the final output included date, permit number, status, tag, address, description, type, and point geometry (point information from address). These were the common features between all of the various data sets.

The last part in the data synthesis process was finding a data storage solution. Having the appropriate data storage solution was important for extraction of information from the data and involved setting up a PostgreSQL database on Amazon Web Services (AWS). The steps to create a local instance of the database was also outlined in a setup guide, so developers can have a self-contained development environment. A PostgreSQL database was used, because the data already takes a relational format, and PostgreSQL provides great extensions for spatial support. This spatial support will enable better analysis on the point features that the building permits present themselves as.

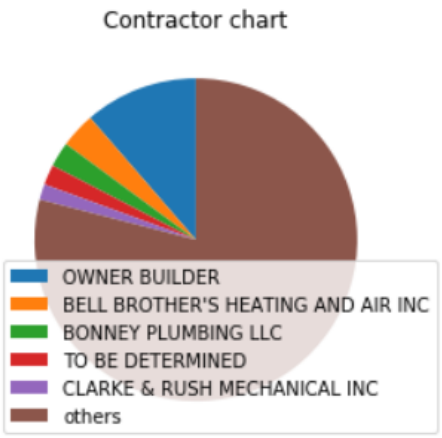
4.2 Data Analysis

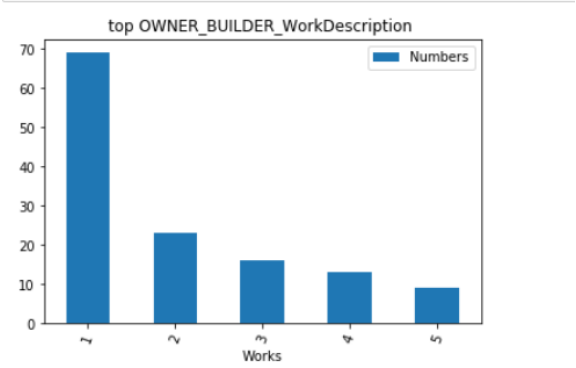
In the era of big data, data analysis is a critical task. Data analysis is the process of discovering valuable and potentially useful information and knowledge hidden in massive, incomplete, noisy, fuzzy, and large databases. It is also a decision support process. Analytics toolchains are based on artificial intelligence, machine learning, pattern learning, statistics, etc. Through the highly automated analysis of big data, inductive reasoning can be used to dig out potential models, which can help companies, businesses, and users to adjust market policies, reduce risks, face the market rationally, and make correct decisions. Among all the analytical tools, we have tried several of them, and finally get to the best one that reacts the best to our data.

Before applying any of these methods, we did some exploratory data analysis (EDA) to see if there's any significant pattern can be found in the data itself. In words, we carried out investigations on the raw data in order to seek for patterns, to spot anomalies, and to check basic statistical assumptions using summary statistics and graphical visualization.

After the EDA process, we'll reform our raw data to the format we need to use for modeling with the help of data processing methods, such as aggregation and filtering. Finally, we'll apply different modeling methods mentioned above to the data and see how each one performs with our data. Our current modeling candidates are listed below. We'll see how they perform as soon as our data is ready.

First off, we implemented some visualization techniques to find patterns and relationship in the building permits datasets. Since the building permit dataset in each county looks similar and it would be too much workload to visualize building permit of all the counties as a whole, we picked the county with the largest number of attributes, Sacramento County, to do a sample analysis. We first used pie chart to plot the top contractors in the dataset. We then used bar chart to illustrate the top work distributions of the first contractor: OWNER BUILDER. These charts might help our sponsor with business decisions. For instance, they can do some investigations to see whether their competitors offer solar energy and electric energy. From the bar chart, our sponsor can further explore the semantic meanings of the work descriptions. They can check the top appearances of work descriptions to see what type of energy the residents prefer to use the most.



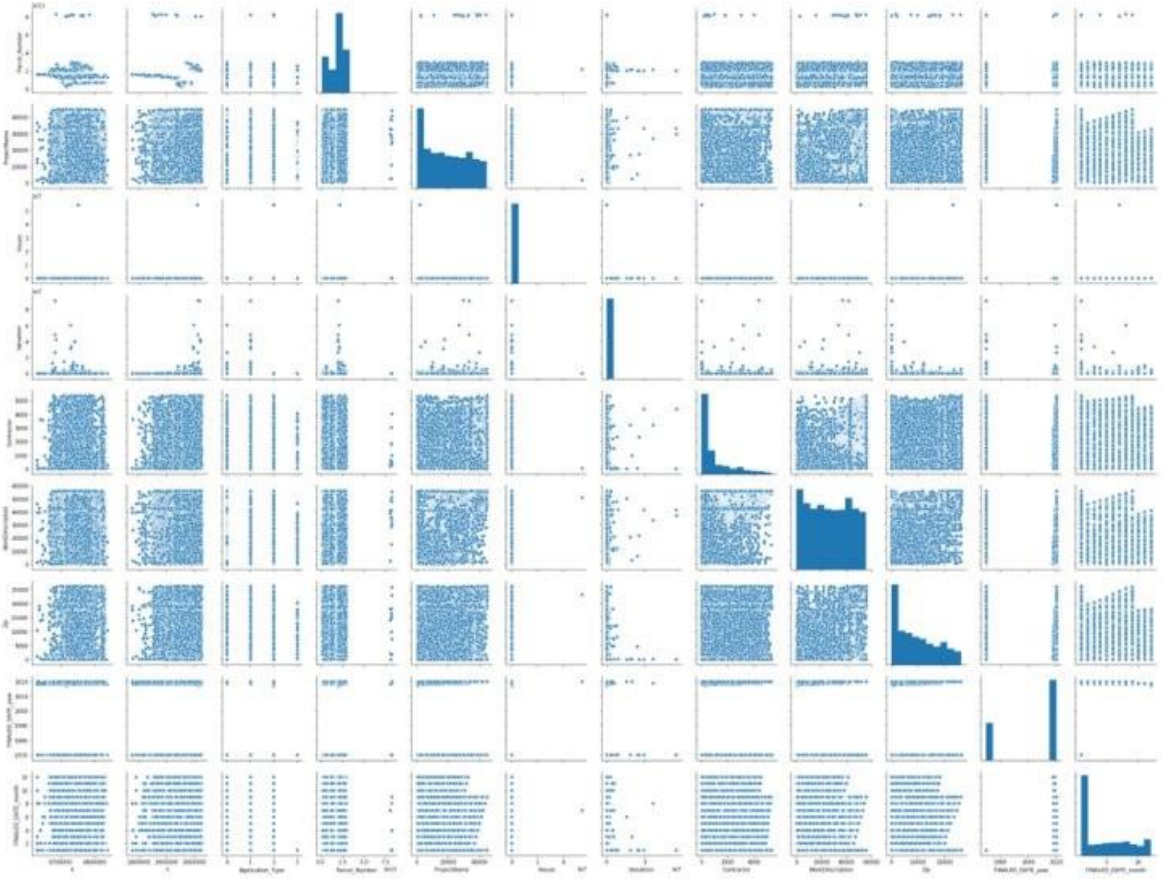


{1: '1. Installation of new, vent less washer/dryer combo in bathroom\n2. Remove and replace all plumbing fixtures\n3, Remove and replace all electrical fixtures, receptacles, switches\n4. Remove and replace kitchen and bath sink and countertops EDR file',
 2: '108, 131, or 132 inspection required prior to covering any work; 108, 131, or 132 inspection required prior to covering any work; Tear off existing asphalt and replace with Class A, Solid, asphalt roofing.\nCarports not included.',
 3: 'Replace existing HVAC condenser whips with new 10/3 whips less than 6ft in length\nReplace existing thermostat conduit with 1/2"" liquid tight conduit\nAdd line set insulation to all refrigerant lines',
 4: 'Electrical service upgrade from 100A to 200A',
 5: 'Clean up, no work is being performed in this space.'}

After these simple visualizations, we selected attributes that might be useful and convert them into numerical ones. We avoid using one hot encoding, which would lead data columns going extremely sparse and running out of computer's memory. We split out attributes FINALED_year and FINALED_month and standardize the large number into small numbers.

	Application_Type	Parcel_Number	ProjectName	House	Contractor	WorkDescription	Zip	FINALED_DATE_year	FINALED_DATE_month
0	0	-1.406148	0	0.0	0	0	0	0.639179	-0.458276
1	0	-0.880818	1	1.0	1	1	1	0.639179	0.091554
2	0	-1.541081	2	0.0	2	2	2	0.639179	0.091554
3	0	0.122392	3	1.0	1	3	3	-1.465526	-1.008106
4	1	-1.748617	4	1.0	3	4	4	-1.465526	-1.008106

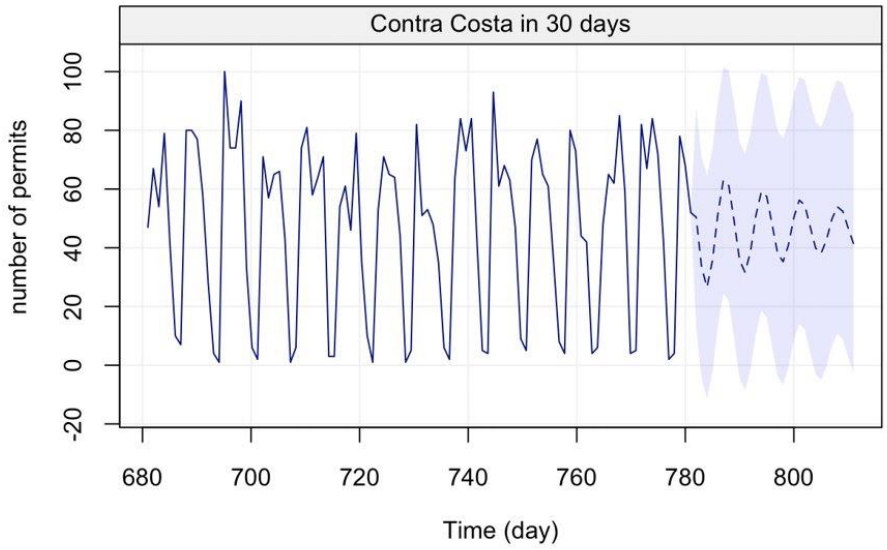
After that, we draw pair charts to check the interaction between all possible pairs.

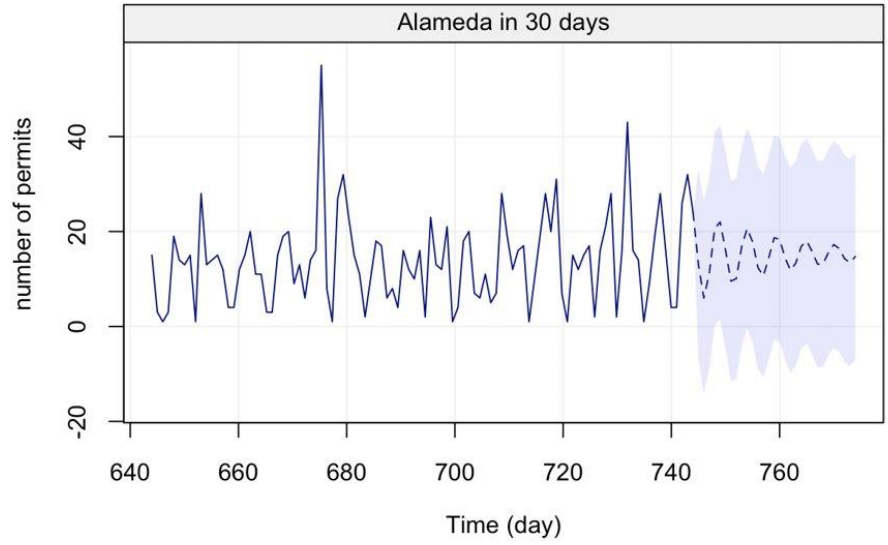


We then applied a DNN model to our cleaned data. Both the pair plots and the modeling result show that there is no relationship between any combination of variables. Since Sacramento County data is the most outstanding and comprehensive dataset we collected, together with this result, we could make a conclusion that the building permits are not correlated with any factor in the dataset.

Since the analysis on the inter-correlation between variables did not work, we tried to think this issue from a time interval aspect. Since it's reasonable to anticipate a relationship between number of permits and seasons, we tried to build a Time Series model for our datasets.

Fortunately, Time Series does work on our data. Besides the largest Sacramento County data, we also choose other two smaller counties to carry out analysis. The following two plots are the prediction of the number of permits in next 30 days in each county. As we can see from the plots, the number of permits fluctuates periodically within a week. Based on such result, our sponsor could change their advertisement targets periodically.

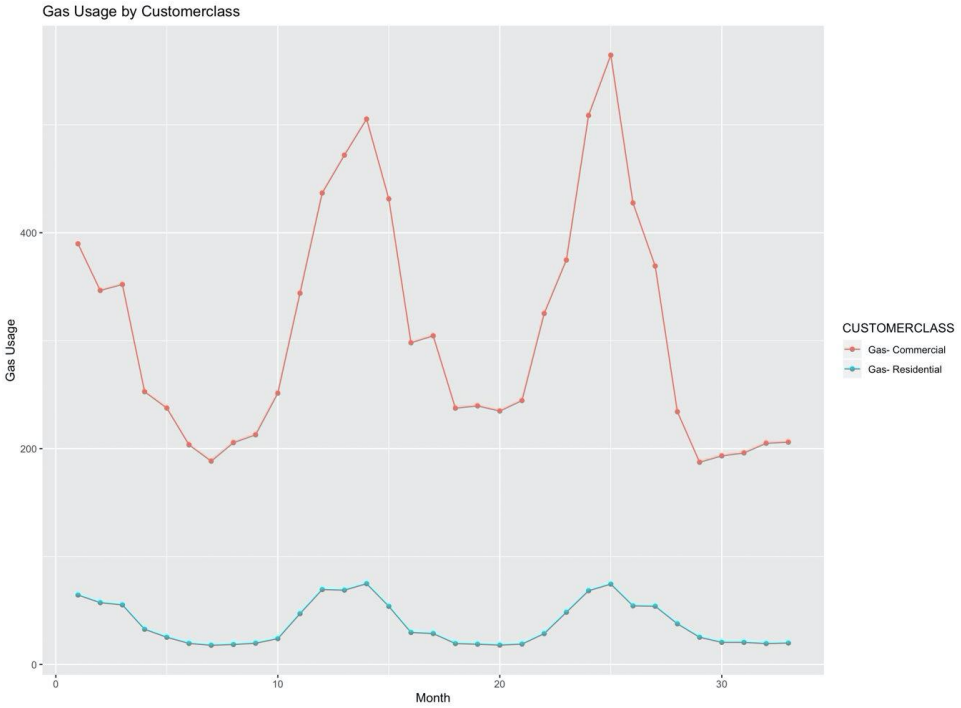


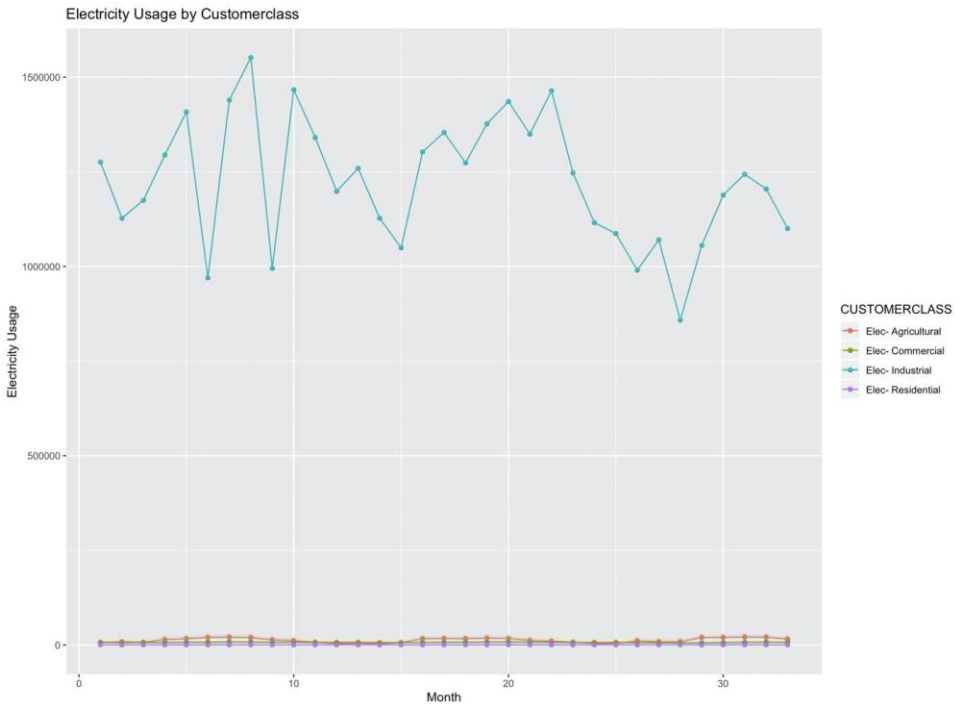


Moreover, besides the building permits analysis, we also looked at the utility usage data and tried to find some useful information that might be helpful to our sponsor. We separate the utility data into gas usage data and electricity usage data, and we explored each of them using Time Series model.

Before applying Time Series model, we firstly do EDA and visualize the data to check statistical assumptions. The first apparent phenomenon is that the electricity and gas in the entire California are mostly used for industrial and commercial purpose. This fact might be helpful to

our sponsor to make a clearer decision on their target customers. Besides the targeting suggestion, by looking at the following two plots, it's interesting to notice that gas and electricity usage follow certain patterns.

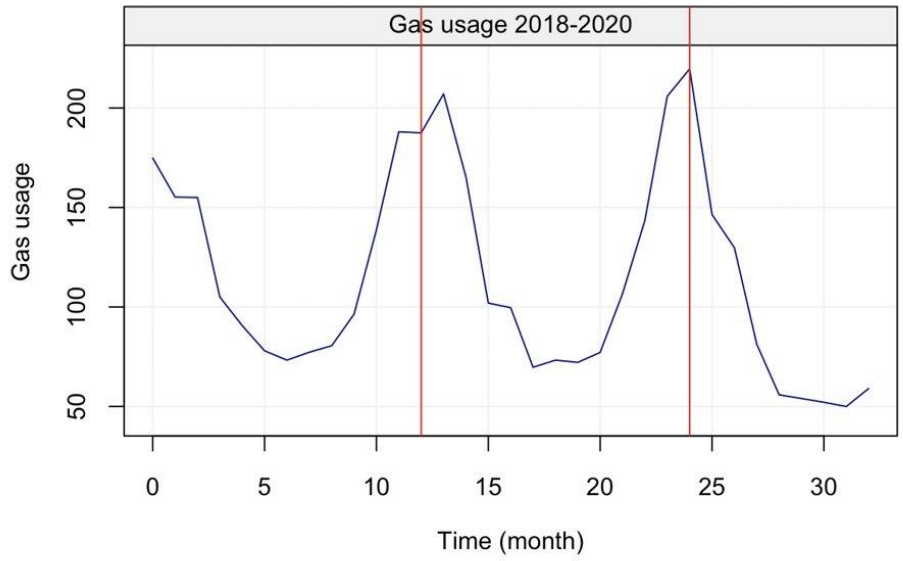


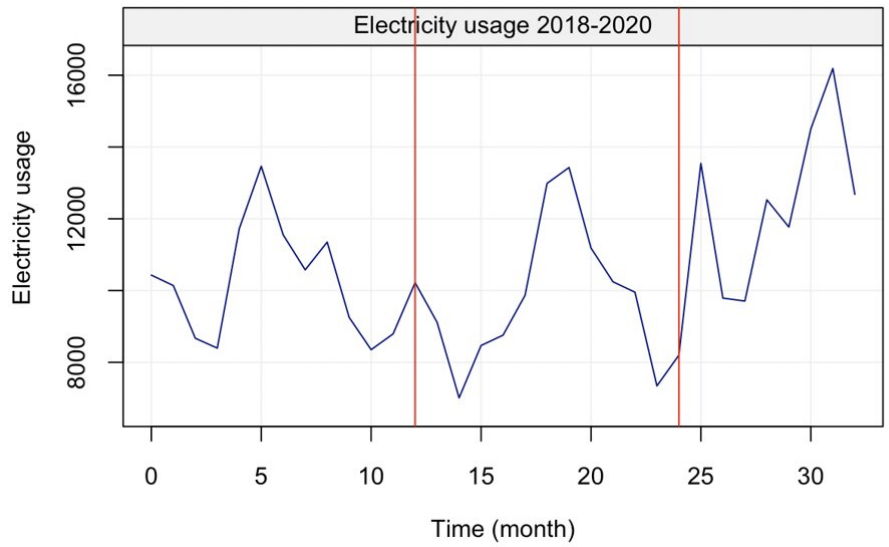


Then we apply corresponding Time Series model to our gas and electricity data separately. Indeed, as we expected, there are patterns existing in gas and electricity usage in each year.

The gas usage is decreasing from spring to summer. It reaches minimum in the mid of summer each year, and then start to increase. It reaches maximum in the end of the year. Then the same pattern shows in the next year.

The electricity usage decreases from January to March. Then it starts to increase from April and reach maximum in the mid of summer each year. Then it starts to decrease again until October. It increases again and reaches another small peak in December. Then the same pattern shows in the next year.



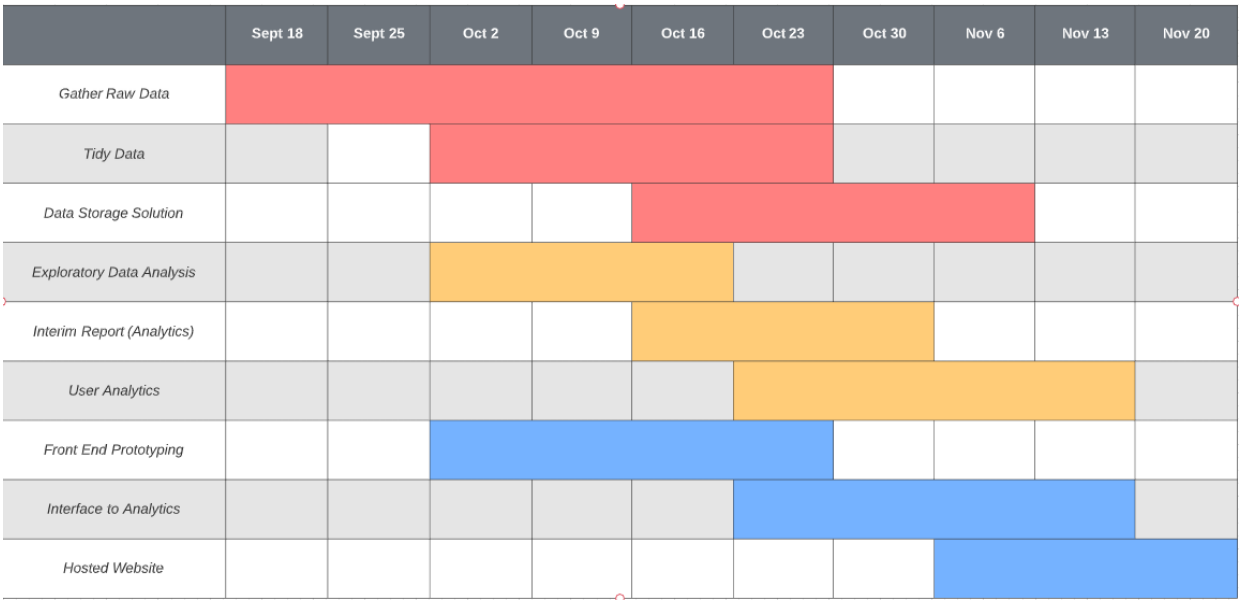


This finding is not surprising, but it's extremely precious for our sponsor. As long as the utility usage trend in one year is predictable, our sponsor could change the business plan to make more profits accordingly.

4.3 Interface Development

The user interface was developed using a WordPress framework. Both static templates for desired pages and a dynamic page featuring a map was developed. Focusing on the map views, this page featuring a map was developed using OpenLayers, a JavaScript library. Using OpenLayers we can query Web Mapping Service (WMS) layers from a server such as GeoServer, which is an open-source server that can plug into our PostgreSQL PostGIS data store. The GeoServer will server tiled images to the map to display and only return information when the on-click function asks for building permit data info.

4.4 Timeline of Work



5.0 BENEFITS

Every team member and our sponsor should benefit from this project. Each team member takes its part to develop the unique feature of this project, ranging from data gathering to front-end design.

Challenges:

The major challenges of this project would be:

1) The time zone difference among our group members. Almost everybody is in different time zone.

2) The project requirements are time-consuming, because our sponsor expects us to gather a relatively huge amount of construction data from 10+ counties with various file types. The data analysis team should work with the data gathering team and will spend more than 50% time cleaning the data. Our back-end team and front-end team need to work together to develop a user interface from scratch.

Efficiency:

By doing this project, our team is expected to develop a webpage/app containing all the desired features so that our sponsor can access the information contained within the data with just a few clicks. For example, whoever accessing our webpage/app should be able to view the construction permits issuing frequency in a specific county since 2016-2020 with projection on how the frequency would change in the next few years.

Performance:

The most important part of this project would be a unique chance providing students with real-world project experience in such a cross-functional team. Students are expected to gain insights on how to work with your co-worker/ manager in their future career paths. So, everyone in the team is expected to develop professionalism on delivering presentation to the sponsor with a nice and niche webpage/app, and to enhance their technical skills.

6.0 DELIVERABLES

The deliverables of the project are listed below. These deliverables follow the categories defined in the approach section and follow from discussions with the sponsor. The due date for the data deliverables was agreed upon between the sponsor and team to allow proper time to perform analyses on the data before the final project report. These deliverables include

- Raw Permit Data
  - Compressed CSV containing building permit data from requested sources.
- Tidy Data file
  - SQL from PostgreSQL data dump that includes tidied data.
- Local Database
  - A hosted relational database containing the tidy version of the permit data.
- Interim Report (Analytics)
  - An interim report detailing exploratory data analysis and propose analytical tools.
- Final Report
  - Details results of the project.
- Hosted Website
  - A hosted website for sponsor to use as a data interface including the code that is run.
- Developer Guide
  - A manual detailing back end and front-end setups.
- Web Site Prototype
  - Prototype of UI/Website for sponsor feedback.
- Final Video
  - Final video that will be used in project presentations.

It is critical that a central data repository is created to store a compiled version of the building permit data in a tidy format. This central data repository should at its core stored as a CSV, since a relational format is appropriate, but the data should be hosted on an instance of PostgreSQL for use with the website front end. By the end of October (23rd) this data warehouse should be completed.

The sponsor will receive all deliverables via email in a compressed archive. This deliverable will encompass the final report, source code of the website, website pages, local development setup

script, developer guide, and the requested data. The application setup includes PostgreSQL, MySQL, WordPress, and GeoServer all of which have setup instructions within the deliverable archive.

Data in this compiled source should include the building permit data from the following counties:

- Sacramento
- San Joaquin
- Contra Costa
- Yolo
- Solano
- Alameda
- Santa Clara
- Santa Cruz
- San Benito
- Monterey
- Napa
- Sonoma
- Marin
- (Optional) San Francisco

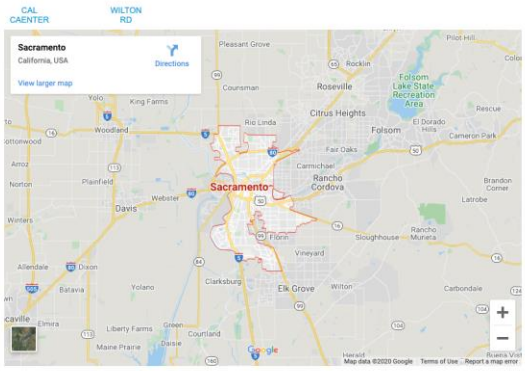
Table 1: Deliverables Table

Deliverable Name	Type (report, prototype, etc.)	Expected Delivery Date
Raw Permit Data	Compressed CSV	October 23, 2020
Hosted Database	CSV/PostgreSQL DB	November 20, 2020
Interim Report (Analytics)	Report	October 30, 2020
Final Report	Report	December 15, 2020
Hosted Website	Code/Website	November 20, 2020
Developer Guide	Manual	November 20, 2020
Web Site Prototype	Code/Website	November 6, 2020
Final Video	Video	December 15, 2020

We have gone and gathered some utility data from PG&E, which we plan on using to help create a model for our sponsor to better understand which homeowners may be potential customers, and how much they might be able to save based on their utility usage, if it is granular enough.

The initial user interface mockup was created and reviewed in a sponsor meeting. This mockup received approval from the sponsor and the front-end team has made a basic prototype based off that mockup. Also based off sponsor feedback on which variables in the data are interesting, we

visualized some of the contractor distributions. We are also able to apply this to other columns in our data.



Backend work has been completed for the prototype. The database is up and running with the PostGIS extension now enabled. Backend team has also setup GeoServer to act as a Web Mapping Service and Web Feature Service that will enable the front-end team to integrate the services into OpenLayers. This will create a front-end website with a map that provides clickable points to display building permit data. The backend team also started work on the technical manual that will be included in the final deliverable so that the sponsor will be able to continue work developing the prototype further after work is done. The details created in the technical manual so far include setting up WordPress and GeoServer using Docker as well as creating a web page with custom script information.

We are making progress on the poster for the showcase. The poster includes background and objectives for the project and discusses the machine learning algorithm we applied on the dataset and impact of our final delivery.

Posters are attached:

DS 440 Public Construction Team

Yixuan Wang, Shunqi Zhang, Joseph Sepich, Myung Joon Kim , Hanzhong Ye
Department of Engineering, Capstone Project supervised by Marc, Rigas and Nate Parkinson

Introduction

Epic Energy

A California-based energy efficiency home improvement organization

- Provide a platform for Epic Energy to make better strategic business decisions and to harness the mass amounts of public data
- Visualize the types of construction projects in different areas

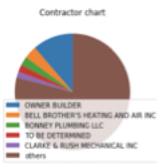
Objective

- Develop a database management system that integrate information from California building permits
- Apply multiple analytical methods to the gathered dataset
- Develop a front-end prototype

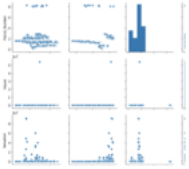
Method 1

- Data mining and cleaning
- Implement data visualization
- Apply DNN model

Data Analysis



Piechart: Group by different builder



pairplot: independency of variable interaction

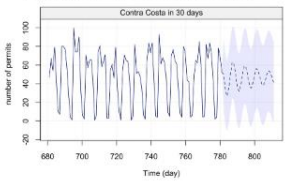
PennState

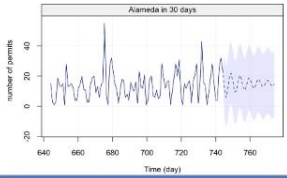
DS 440 Public Construction Team

Yixuan Wang, Shunqi Zhang, Joseph Sepich, Myung Joon Kim , Hanzhong Ye
Department of Engineering, Capstone Project supervised by Marc, Rigas and Nate Parkinson

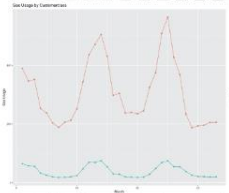
Methods

- Data mining and cleaning
- Implement data visualization
- Apply Time Series model
- Predict the amount of permits in next 30 days of each county

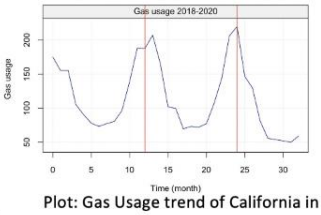




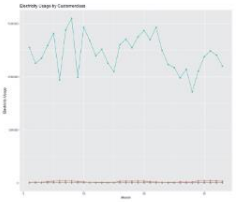
Data Analysis



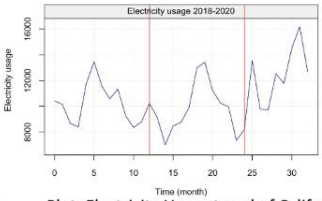
Dotplot: Gas Usage trend of different type of users in California



Plot: Gas Usage trend of California in past two years

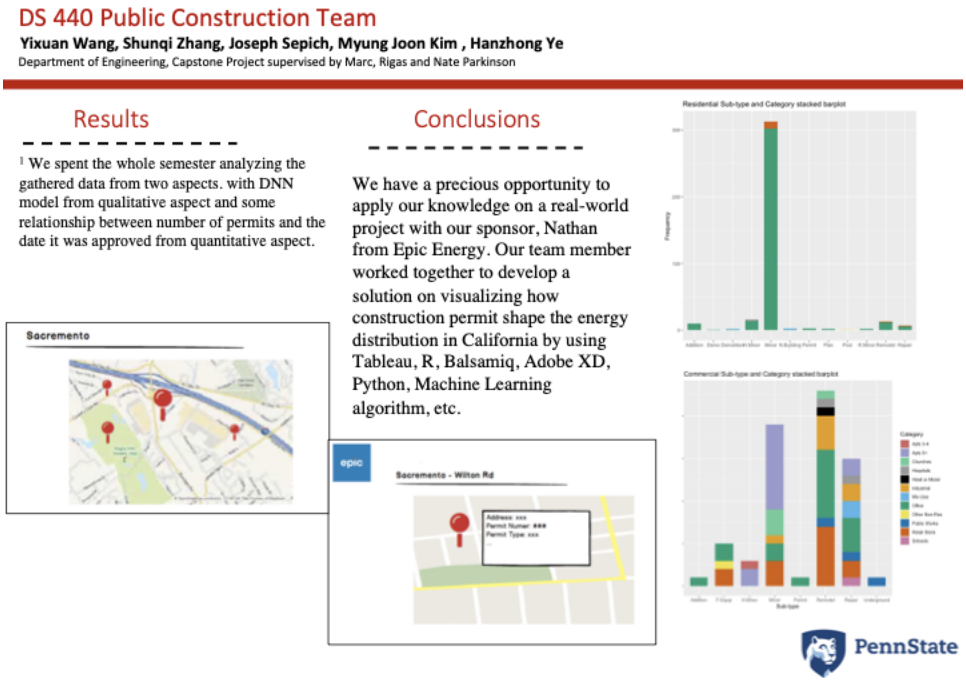


Dotplot: Electricity Usage trend of different type of users in California



Plot: Electricity Usage trend of California in past two years

PennState College of Engineering



7.0 TEAM CAPABILITIES

Joseph Sepich has experience in data engineering including data wrangling and data storage solutions. These data engineering skills will play into his role as a data gatherer. Knowledge using R and Python with enable him to create scripts to transform raw data from various data sources into a single tidy data format. PostgreSQL experience will help the team to insert the data into the database and fully use the spatial aspects to query the data for analysis.

Hanzhong Ye has experience in data analyzing including machine learning and deep learning. These data analyzing skills will play into his role as a data analyzer. Knowledge using PyTorch with enable him to create neural networks and data models, matplotlib to visualize data. SQL and NO-SQL experience will help the team to insert the data into the database and fully use the spatial aspects to query the data for analysis.

Yixuan Wang has experience in data analyzing and transforming datatypes using Python. These data analyzing skills will play into his role as a data analyzer and raw data processor. Knowledge using Python packages (pandas, NumPy and matplotlib) will enable him to data-mining on raw data set and R to visualize data by developing data models. Also, knowledge using Balsamiq and Figma would enable him to offer some help to UI/UX designers of this project.

Manasvi Mittal has experience in using HTML and CSS for front end designing. He also has experience with python and coding. He has CAD modeling skills too. This variety of knowledge can come handy with data visualization and analysis. He currently is learning R which can be helpful in data visualization as well.

Shunqi Zhang has experience in data analyzing and data mining using R and Python. These data analyzing skills will play into his role as a data processor and data analyst. Knowledge using R and Python will enable him to generate machine learning methods and network methods. Knowledge using R and Matlab will enable him to do visualization before constructing models.

Myung Joon Kim has experience in using HTML and CSS for front end designing. He has experience in making website prototypes as well. This front-end skill will be usefule for his role as a front-end developer. He also has experience in combining back-end with MySQL and Flask, a web framework in Python. This experience would enable him to help data organizing for the team.

8.0 BUDGET NARRATIVE

The only expected costs associated with this project were purchasing cloud services for hosting data and/or the web interface. According to the Simple Monthly Calculator (https://calculator.s3.amazonaws.com/index.html) provided by AWS, a single EC2 instance (used 3 hours/day) with associated storage (around 80GB) is estimated to cost around $65/month. On a three-month range for the current project that would bring the total to around $200. These costs did not end up being used as PSU would not approve the purchasing of the software for the project; however, this gives insight on how the maintenance costs of the project would be in a productive environment.

APPENDIX B: PROJECT BUDGET

Project Start: September 18, 2020

Project End: November 20, 2020

Budget Item	**Proposed**	**Interim**	**Actual**
Travel	None - 0	0	0
Equipment	AWS - $200	0	0
Materials	None - 0	0	0
Miscellaneous • Project poster • Others…	None - 0	0	0
Total Project Costs	**200**	**0**	**0**