Analysis of Energy Consumption by various departments in San Jose

DATA 225 : Database Systems for Analytics

Submitted by (Group 6) :

Deekshita Prakash Savanur (016597815)

Gouri Benni (016285698)

Renita Dsilva (016068312)

Sawan Shivanand Beli (016522662)

Abstract - To encourage the recent rise in worldwide awareness of the need to reduce energy use, many measures have been put in place. The entire public now has access to energy use data for research and finding ways to save. It can be difficult to understand anything about energy use since, as climate change and global warming worsen, more factors come into play. Keeping this in mind, we want to examine a sample of energy usage statistics from several San Jose city agencies. Data understanding and gathering, data processing to transform the data into a format suitable for data analysis by removing any outliers and erroneous data, data analytics, and the creation of an analytical report for the process are the procedures involved.

Index Terms - Data Architecture, Data Models, Datasets, Data Visualization, Data Warehouse, Data Analysis.

I. Motivation

Since we have recently moved to San Jose and have begun independent lives there, where we deal with numerous bills each month, we have come across the PG&E bills that tell us how much energy we have used during a specific month. It compelled us to learn more about the energy used by various departments in San Jose city because we were confident that different buildings and their departments would have varying consumption bills.

We were curious to find out how much energy each department used in a given year so that we could perform statistical analysis and visualisations based on various scenarios and predict what the consumption rate of the various departments in the city would be. We are attempting to move toward green remodeling and lower energy usage in the future through the analysis. Here, we'll use an energy consumption dataset that displays information about the energy used by several San Jose departments, including aviation, parks, human services, housing, public transit, conventions, fire, and libraries, and run an analysis to draw out pertinent data.

II. Literature Review

The trends in energy consumption over time have been recognized and reviewed by several studies and research projects. To get inspiration for our study, we have cited a few of the numerous research articles that have been written on this subject.

- Using openly available public data, this paper suggests a "Big Data Analysis Process for Residential Housing Energy Consumption" written by W Pak, Inhan Kim, and Jungsik Choi. The four stages of this process are as follows: Understanding data, including investigating and gathering information on architecture, weather, and energy use; Data transformation involving the conversion of residential energy consumption data and reference input data into master data, which is analytical data that has been processed through the filtering, refining, and type conversion of the acquired data, for big data analysis; Evaluation, data evaluation, and application of the analytical model are all part of data analytics, which is the development and deployment of an analysis model for the energy consumption of residential buildings.
- The authors Konstantin Hopf, and Mariya A. in their paper "Energy Data Analytics for Improved Residential Service Quality and Energy Efficiency"

demonstrated how to make use of the little data generated by active online users to deliver low-cost and extensive insights to possibly all residential utility consumers. They do this by utilizing Green IT artifacts built on machine learning that enhance decision-making, the efficacy of energy audits, and conservation campaigns, thus raising customer value and encouraging the uptake of associated services. Furthermore, demonstrated how decision quality can be significantly increased by using information from publicly accessible geographic information systems.

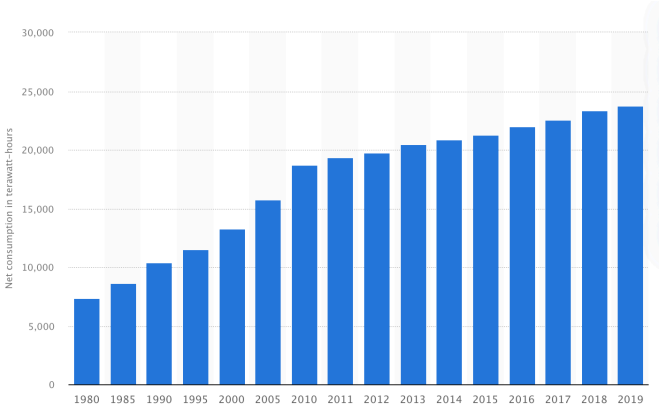
• "Applying Big Data Analytics for Energy Efficiency" by Hussnain Ahmed focusses on seasonal and daily usage trends by analyzing the data. The classification of buildings based on their energy efficiency is also included in the analysis, and the effects of the seasons are also taken into account. A model for separating energy-efficient buildings from those that are not was created using the analysis.

• "A three-year dataset supporting research on building energy management and occupancy analytics" - This study by authors Na Luo, Zhe Wang, David Blum, Christopher Weyandt, Norman Bourassa, Mary Ann Piette & Tianzhen Hong provides the curation of a monitored dataset from a Berkeley, California, office building that was built in 2015. The dataset comprises person counts, HVAC system operating conditions, indoor and outdoor environmental characteristics, and whole-building and end-use energy consumption. Over 300 sensors and meters were used to collect the data over three years from two office levels (2,325 m2) of the building.

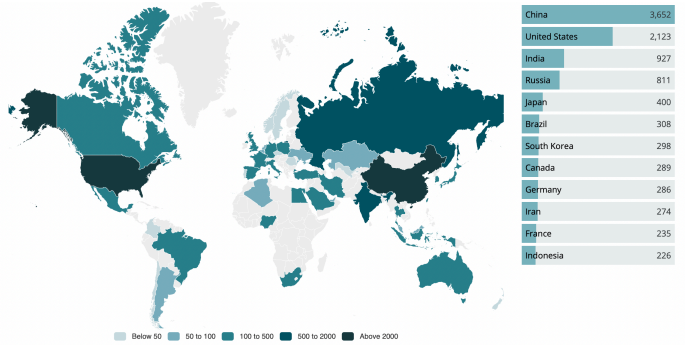
III. World Statistics

The following data presents a brief understanding of the power consumption statistics in different parts of the world.

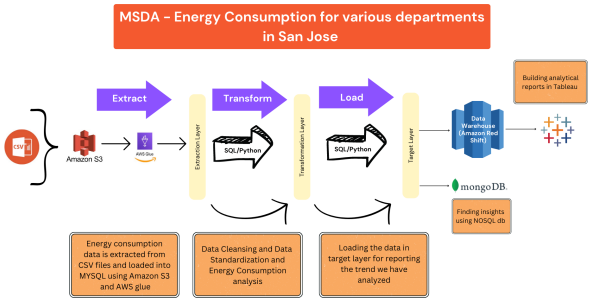
1. Over the previous 50 years, the amount of energy consumed worldwide has increased steadily, reaching over 23,800 terawatt-hours in 2019. The use of energy more than quadrupled between 1980 and 2019, while the world's population grew by almost 75\%. Global industrialization expansion and increased access to energy have both increased global electricity demand.

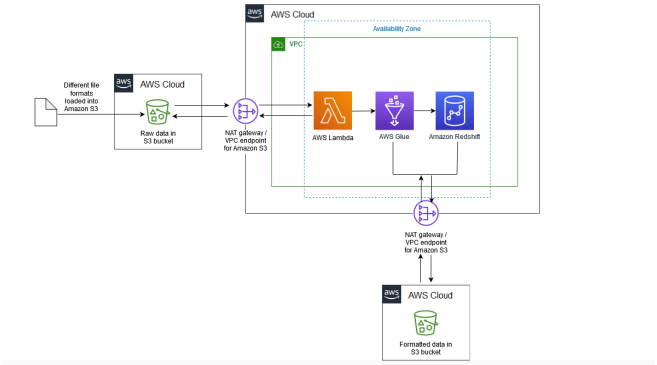


2. Around 87 percent of the world's population, according to the International Energy Agency, an OECD intergovernmental institution, has access to electricity. Less than three-quarters of the world's population had regular access to electricity in 2000, according to estimates from the IEA. For the first time, less than a billion people worldwide do not have access to electricity, with that percentage currently just shy of 90 percent. The majority of those without electricity live in sub-Saharan Africa, one of the world's poorest regions. While the majority of people in many nations still live without electricity, case studies conducted over the past 20 years have demonstrated that nations may increase access to electricity for the vast majority of their citizens in less than ten years.



IV. Architecture Diagram





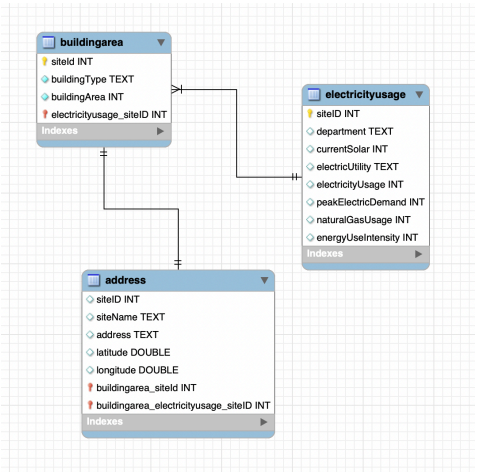
The above diagram represents the workflow in the AWS Cloud Platform.

V. DESIGN STEPS

Description	Tools/process used
Scope of Project	Google Meet, Group Thinking
Project Proposal	MAC Pages, Google Meet
Dataset Study	Data.gov,Kaggle,Other websites
Modelling of Data	draw.io
Dataset View	Excel(.csv)
Analysing data	MySQL Workbench
Data Mapping	Jupyter notebook,Python lib.
Data Documentation	MAC Pages
Development(SQL)	MySQL Workbench queries
Development(NoSQL)	MongoDB queries
Development(AWS Cloud)	Amazon s3, AWS Glue, VPC endpoints
Data Warehouse	AWS Redshift
Data Visuals	Tableau,Python matplotlib
Project PPT	Prezi
Project Report	Overleaf(Latex)
Version Control	GitHub

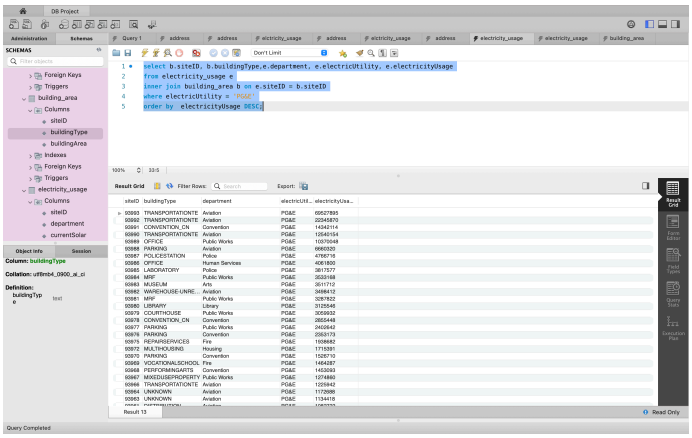
VI. ER DIAGRAM

By reverse engineering method in MySQL Workbench, we have derived the following ER diagram.

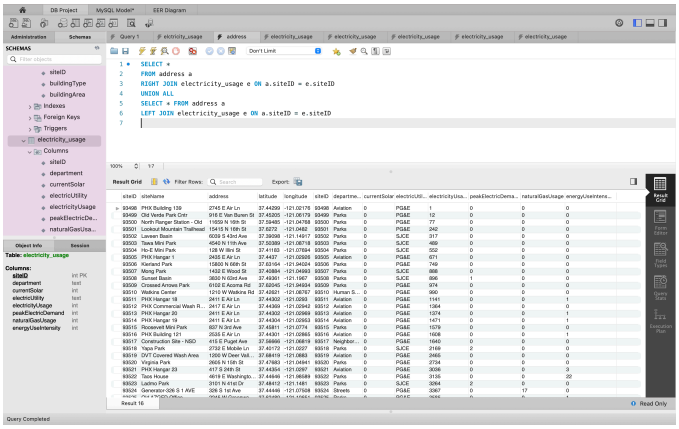


VII. ENERGY CONSUMPTION ALGORITHM

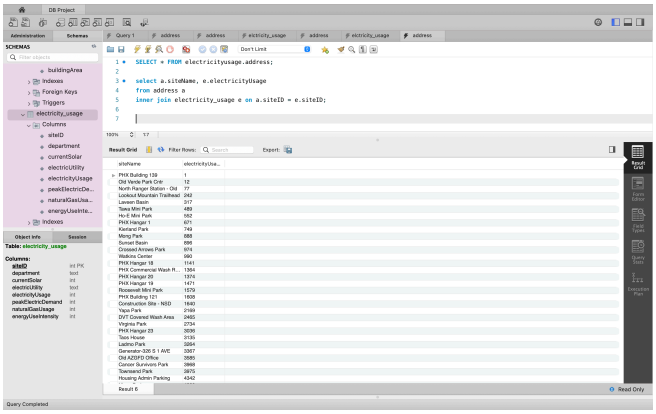
1. The following query results in the electricity usage in various departments of the city which are under PG and E electric utility companies arranged according to the highest to lowest consumption rate. It shows the columns siteID, building type, department, electric utility, and electricity usage from two tables electricity_usage and building_area. The query runs a select statement that displays attributes siteID, building type, department, electricUtility, and electricity usage by using an inner join on two tables Electricity_usage and building area on a common column siteID where the electric utility company is equal to PG&E and is ordered in a descending manner of electricity usage in various departments having numerous building types with Aviation having the highest electricity usage.



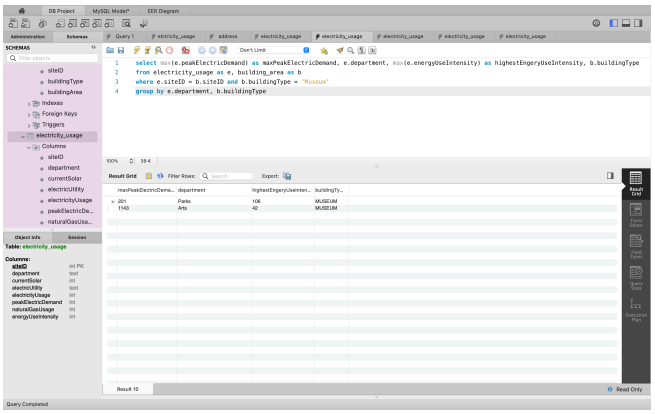
2. The query results in the summation of both tables to analyze the data and has 2 part where the first select statement returns records from the right table electricity_usage with a join condition where all attribute values will be displayed and the second select statement return records from the left table address. Both the results are displayed together using ‘union all’.



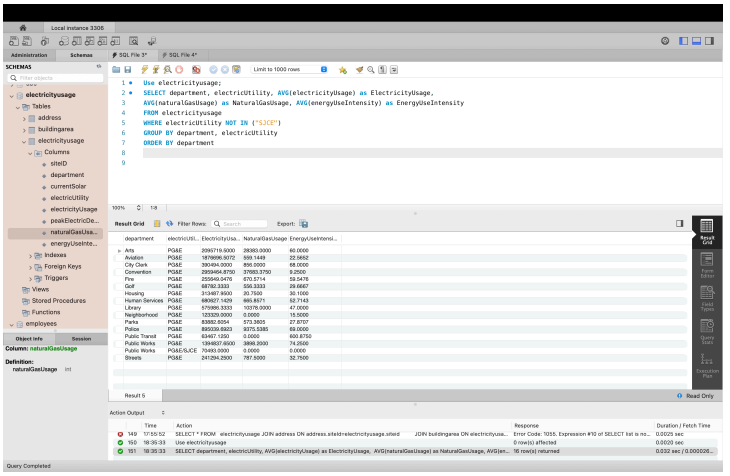
3. Here we get results for the amount of electricity consumed by various sites in San Jose through joining two tables i.e. the address and electricity_usage using inner join on column siteID with attributes shown i.e siteName and electricity usage by that particular site. The values are arranged in ascending order with the least usage site name along with the usage rate at the top of the findings and the highest usage site name along with the usage rate at the bottom of the sheet.



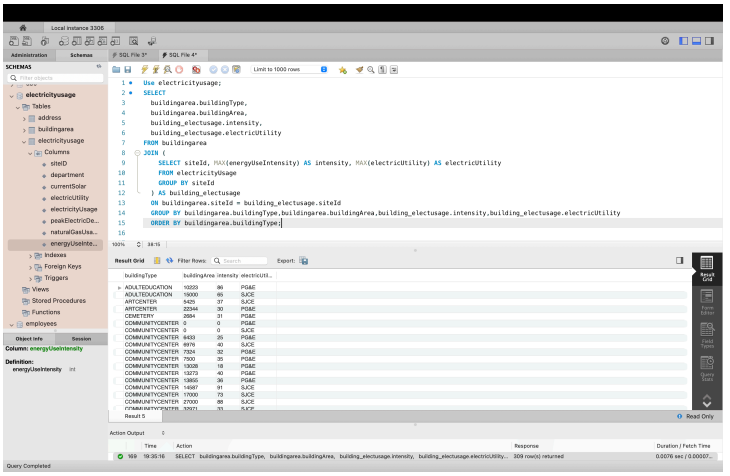
4. The below query results in showing the maximum Peak Electric Demand and highest energy use intensity for a particular building type Museum the data is grouped concerning the attributes department and building type and we conclude that the Museum type building has two departments that are Parks and Art having the highest consumption rate. The query selects the attributes such as department and building type and uses aggregate function ‘max’ to find the highest electric demand and highest energy use intensity from two tables electricity usage and building area having two where conditions having an operator to provide an output that satisfies both conditions of having the same siteID and the building type must be a Museum and grouped according to the department and building type.



5. The query results in providing an average usage of attributes like electricity usage, natural gas usage, and energy use intensity within various departments in San Jose and which results in data belonging to one utility company that is PG&E. The select statement returns attributes related to the department, electric utility company, averages of electricity usage, natural gas usage, and energy use intensity from electricity usage from the electricity usage table with a where condition that returns data related to utility companies except SJCE and grouped based on department and electric utility attributes and ordered alphabetically based on department column.

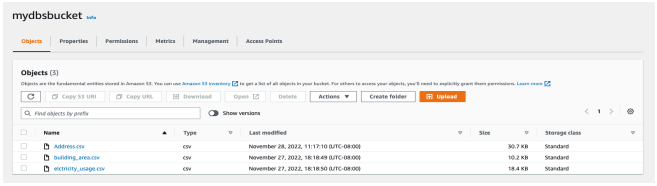


6. The below query results in maximum energy use intensity and most preferred electric utility company along with the area covered by a particular building type and the data is grouped based on the building type, building area, energy use intensity, and electric utility company, and ordered in an alphabetical order based on the building type attribute.

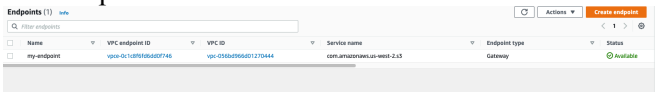


In AWS Cloud, the following steps were followed :

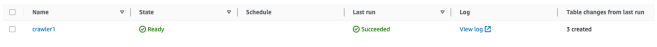
- Creating an S3 bucket and uploading data files.



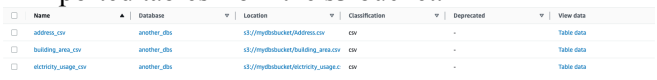
- Creating VPC endpoints for the ETL process of the data present in an S3 bucket.



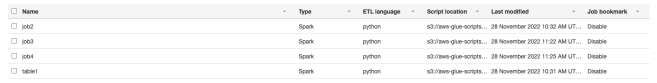
- Creating a crawler to import data from the S3 bucket.

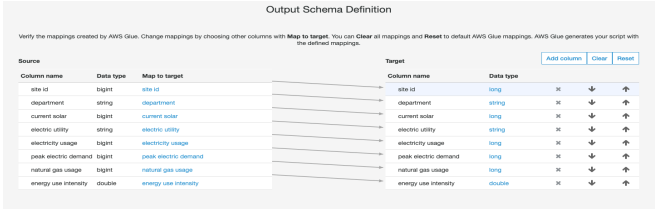


- Imported tables from the s3 bucket.

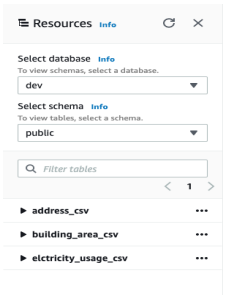


- Creating jobs for each of the tables in the AWS glue database where the data source is the S3 bucket and the data target is the redshift database using the Amazon redshift connection.





- Creating a Redshift cluster for querying.



Few queries :

- Considering 0.35$ per kWh. Find the top 10 building types with the highest cost.

```
>>  select building."building type",
avg(elec_usage."electricity usage") as
electricity_usage, 0.35* electricity_usage as total_cost
   from the public.building_area_csv as building
   right join
         public.elctricity_usage_csv as elec_usage
     on building."site id"=elec_usage."site id"
   group by building.” building type"
   having total_cost > 500
   order by total_cost desc limit 10;
   order by avg_cost;
```

buildingtype	electricity_usage	total_cost
CONVENTION_CN	8598781	3009573.35
TRANSPORTATIONTE	5355298	1874354.30
LABORATORY	3817577	1336151.95
MRF	3410495	1193673.25
COURTHOUSE	3059932	1070976.20
PARKING	881190	308416.50
DISTRIBUTION	805210	281823.50
PERFORMINGARTS	780194	273067.90
VOCATIONALSCHOOL	729252	255238.20
OFFICE	715092	250282.20

- Top 10 departments that use natural gas under PG&E electric utility.

```
>> select elec_usage."department",
avg(elec_usage."natural gas usage") as
natural_gas_usage from public.elctricity_usage_csv as
elec_usage
   where elec_usage."electric utility"='PG&E'
   group by elec_usage."department"
   order by natural_gas_usage desc limit 10;
```

department	natural_gas_usage
Convention	37683
Arts	28383
Library	10378
Police	9375
Public Works	3898
City Clerk	856
Streets	787
Fire	670
Human Services	665
Parks	573

- Top 10 departments that use natural gas under SJCE electric utility.

```
>> select elec_usage."department",
avg(elec_usage."natural gas usage") as natural_gas_usage
from public.elctricity_usage_csv as elec_usage
    where elec_usage."electric utility"='SJCE'
    group by elec_usage."department"
    order by natural_gas_usage desc limit 10;
```

department	natural_gas_usage
Housing	5280
Neighborhood	2974
Police	1537
Fire	935
Public Works	377
Human Services	258
Library	180
Parks	172
Streets	0
Public Transit	0

- Sites with the highest electricity usage are under SJCE utility.

```
>> select elec_usage."department",
avg(elec_usage."natural gas usage") as natural_gas_usage
from public.elctricity_usage_csv as elec_usage
    where elec_usage."electric utility"='SJCE'
    group by elec_usage."department"
    order by natural_gas_usage desc limit 10;
```

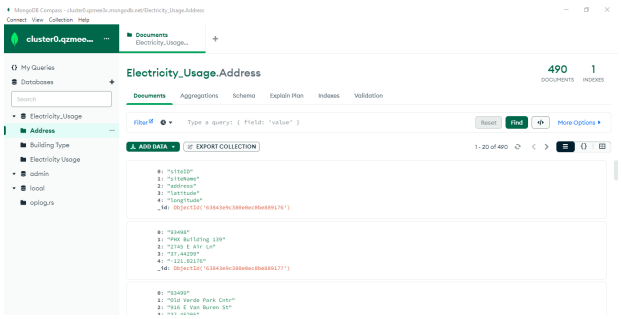
sitename	buildingarea	electricityusage	cost
Pecos Park - Community Center	93670	1595744	558510.40
Steele Indian School Restroom	67470	401605	140561.75
Deer Valley Pool	60000	121351	42472.85
Goelet A Beuf Community Center	48000	728857	255099.95
Encanto Maintenance and Warehouse	42090	250259	87590.65
Paradise Valley Community Cntr	35506	332983	116544.05
Deer Valley Community Center	34294	375779	131522.65
Washington Activity Center	32971	346891	121411.85
Papago Sports Complex	31426	193986	67895.10
Maryvale Community Center	27000	760600	266210.00

- Department and respective building types that are highly powered by solar energy.

```
>> select electricity.” department", building.” building
type”, sum(electricity." current solar") as
total_current_solar_usage, count(electricity.” current solar")
as count, avg(electricity.” current solar") as
avg_current_solar_usage from the
public.elctricity_usage_csv as electricity
    inner join
        public.building_area_csv as building
        on building."site id"=electricity."site id"
    group by  electricity.” department", building.” building
type"
    order by total_current_solar_usage  desc
    limit 10;
```

department	buildingtype	total_current_solar_usage	count	avg_current_solar_usage
Aviation	PARKING	5400	8	675
Public Works	PARKING	1326	2	663
Aviation	OFFICE	580	7	82
Public Works	REPAIRSERVICES	540	4	135
Housing	MULTIHOUSING	461	19	24
Public Transit	TRANSPORTATIONTE	207	14	14
Library	LIBRARY	150	15	10
Parks	COMMUNITYCENTER	140	17	8
Public Works	OFFICE	100	8	12
Convention	CONVENTION_CN	100	2	50

- For the MongoDB part, we have three tables, and all three tables namely Address, Electricity usage, and Building Area are imported to MongoDB. The database name is Electricity_usage.



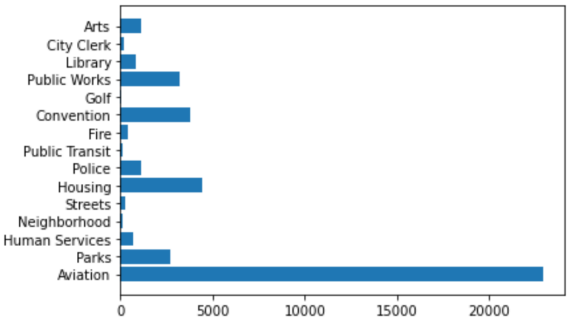
VIII. Data Visualization

The process of converting unprocessed data into useful graphical representations including charts, graphs, photos, and films is known as data visualization. It is helpful to obtain insights from it in this way since it makes it simple to explain the digits and numbers. The relevant images are produced based on the energy consumption algorithm previously discussed.

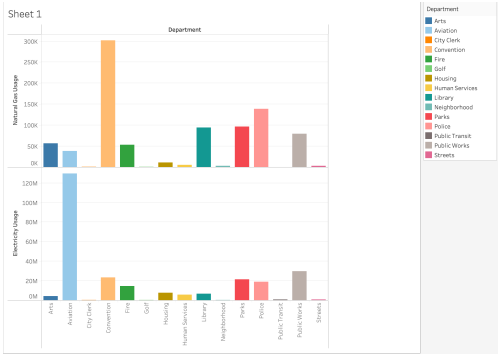
IX. Tableau

Tableau is the program we used to visualize the data. Tableau has a reputation for producing interactive representations that can be tailored to the intended audience.

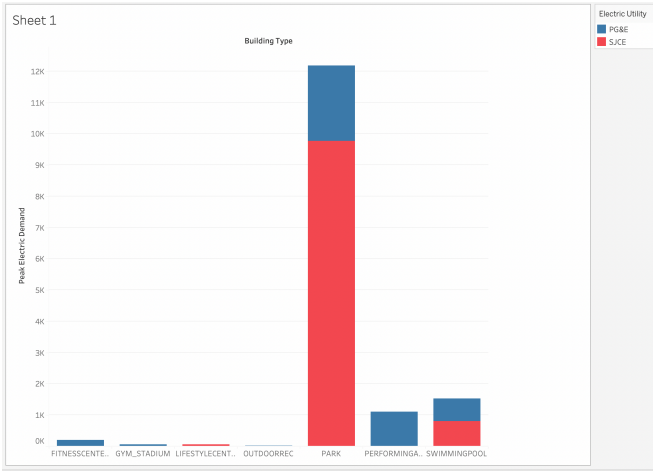
1. It depicts the power consumption worldwide for the year 2021.



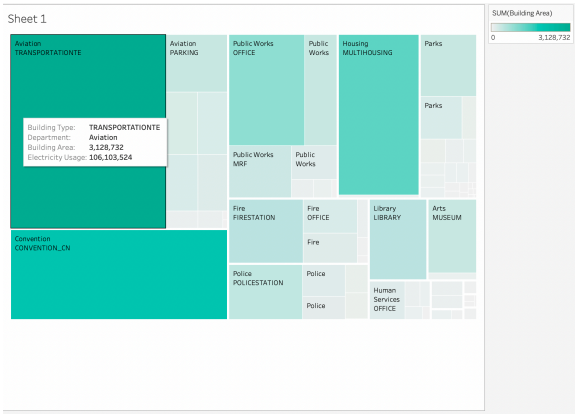
2. The bar graph shows a comparison between electricity usage and natural gas usage for various departments in San Jose.



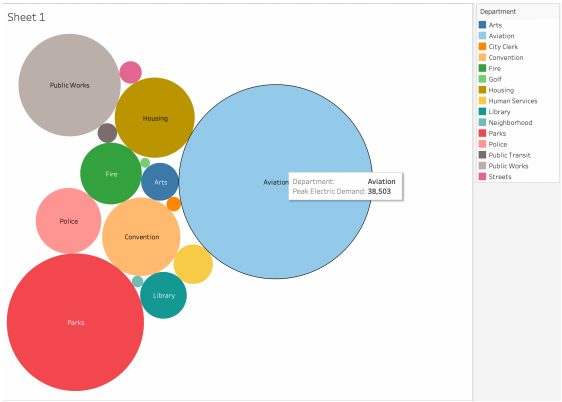
3. A stacked bar graph here visualizes the different building types filtered according to fitness centers within San Jose and their Peak Electric Demand for PG&E and SJCE electric utility companies.



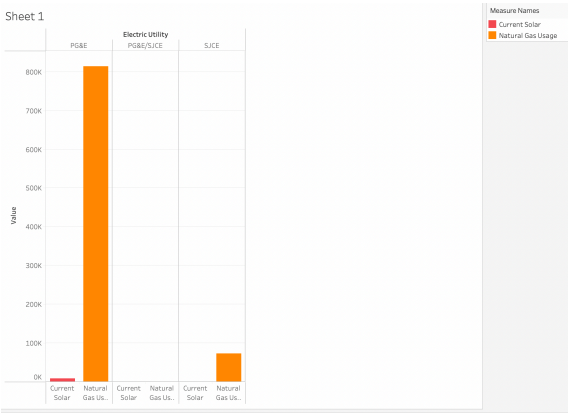
4. Here the heatmap shows the electricity usage for a department having a particular building type with the area of the building and as we see here the highest intensity of color depicts the largest building area which directly corresponds to consuming the highest electricity.



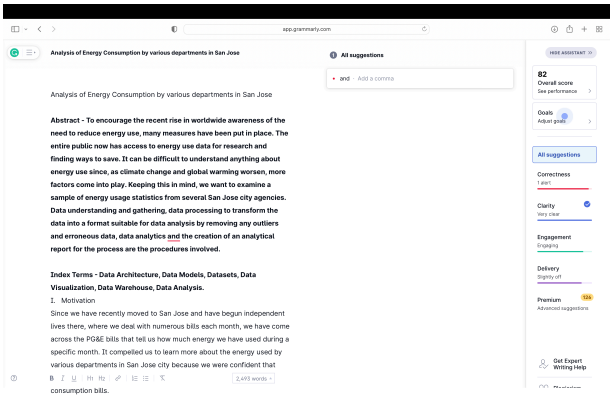
5. As shown in this packed bubble graph we have data indicating the Peak Electric demand for various departments like Aviation, Police, Parks, etc distinguished in different sizes and colors.



6. This side-by-side bar graph shows the current solar and Natural gas usage for various electric utility variants like PG&E, PG&E/SJCE, and SJCE.



>> Use of Grammarly :



X. Key findings

1. While comparing the utility companies PG&E and SJCE we find that natural gas usage is consumed by 95\% of the city in parity with current solar for PG&E company.
2. The highest electric demand is by the Aviation department which is around 38,503 and the lowest electric demand is by the Golf department which is around 99.
3. We found that the electricity usage is directly proportional to building area grouped based on the type of building and department and found that for Aviation with Transportation building has the highest electricity usage and aviation with an unknown building have the lowest consumption rate and similar findings are there for the rest of the departments.
4. Natural gas is highly consumed in the convention department and least in the public transit department while Electricity is highly consumed by the Aviation department and least by the Golf department.

XI. Challenges

1. Establishing distinct tables for each entity/theme by normalizing the denormalized extracted raw data.
2. As we could join two tables in relational database it is not the same case in MongoDb. Joining the stored documents was a tricky task. Since we can manually code it hence it is not impossible but it was time consuming which was eventually affecting the performance.
3. An Amazon Redshift cluster's purchasing price is influenced by a variety of factors. To prevent unpleasant surprises in the future, everyone thinking about using Amazon Redshift as their data warehouse needs to fully comprehend these issues. An in-depth article on Amazon Redshift pricing may be found here. Amazon Redshift caters to all businesses, regardless of size, with a wide range of pricing options and deployment flexibility.

XII. Conclusion

This study proposed the energy consumption data analysis process and analysis model from the definition of the data to be collected to the use of the analysis data for the analysis of the energy consumption data of existing departments in San Jose.By confirming the association between the energy consumption of various departments in San Jose and the architectural aspects such as the area of the building that affect it, the data analysis model for energy consumption in various departments in San Jose city may be used to create energy performance improvement factors for energy savings in the city. In the real world, this trend analysis is used to work towards building green operations and towards using energy in an efficient way which in turn will help in improving the city's overall energy consumption rate.

XIII. Learnings from Project

- Along with the tools discussed in the lectures, the team also experimented with and learned how to use new ones, such as Tableau, Python packages, MongoDB, AWS Cloud, and Prezi.
- The group also looked into a lot of trends in energy consumption by various departments, as well as the comparisons.
- The team gained knowledge of teamwork, agile practices, pair programming, minute-taking, and following through on action items.
- The team also completed some project-related studies on data warehousing and created a prototype.

XIV. References

1. https://www.aceee.org/files/proceedings/1992/data/papers/SS92_Panel10_Paper17.pdf
2. https://www.osti.gov/servlets/purl/1372902
3. https://data.world/
4. https://am.jpmorgan.com/content/dam/jpm-am-aem/global/en/insights/eye-on-the-market/2022-eotm-energy-paper.pdf
5. https://www.ericsson.com/en/reports-and-papers/research-papers/global-electricity-usage-of-ict-network-operators---an-extensive-data-set

XV. Appendix/ Rubrik Explanation

Criteria	Pts	Comments
Presentation Skills Includes time management	5 pts	
Code Walkthrough	3 pts	The presentation's Tools and Architecture segment covers languages, connectors, and other tools. All the scripts and code are uploaded to GitHub, and the project demonstration documentation includes screenshots of every result.
Discussion / Q&A	4 pts	
Demo	5 pts	
Version Control Use of Git / GitHub or equivalent; must be publicly accessible	3 pts	https://github.com/Saavnbeli/Database-Systems-
Significance to the real world	5 pts	To encourage the recent rise in worldwide awareness of the need to reduce energy use, many measures have been put in place. The entire public now has access to energy use data for research and finding ways to save. It can be difficult to understand anything about energy use since, as climate change and global warming worsen, more factors come into play. Keeping this in mind, we want to examine a sample of energy usage statistics from several San Jose city agencies.
Lessons learned Included in the report and presentation? How substantial and unique are they?	5 pts	The report includes all significant sections.

Innovation	5 pts	Bases on the energy consumption analysis insights on various departments in San Jose, we can conclude that we can use green energy usage which in turn lead us to green environment.
Teamwork	5 pts	As a team we practiced pair programming, Agile framework etc..
Technical difficulty	4 pts	As a team we were unaware and learnt querying in MongoDB and Tableau.
This criterion is linked to a learning outcome Practiced pair programming? See: https://en.wikipedia.org/wiki/Pair_programmingLinksLinksLinksLinksLinksLinksLinksLinksLinksLinksLinksLinksLinks Links to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site.	2 pts	A document of Minutes Of Meeting (MOM) is provided as an evidence or proof for pair programming.
Practiced agile / scrum (1-week sprints)? Submit evidence on Canvas - meeting minutes, other artifacts	3 pts	Submitting the document for Minutes of Meeting (MOM) for the project schedule and discussions.
Used Grammarly / other tools for language? Grammarly free version is sufficient; can use other tools as well. Submit report screenshot on Canvas.	2 pts	Grammarly and Google docs were used as other tools and screenshot of each is submitted.
Slides	5 pts	Have attached it in Canvas.
Report Format, completeness, language, plagiarism,	5 pts	Used IEEE format document in Google Docs. For slides, we have used selective features in

whether turnItIn could process it (no unnecessary screenshots), etc		Microsoft Powerpoint for PPTs.
Used unique tools E.g.: LaTeX for writing reports (submit .tex that is not generated from another format such as .docx; generating from .lyx and similar LaTeX editor outputs is fine. Also checkout https://www.overleaf.com/LinksLinksLinks Links to an external site. to an external site.) Unique features of Prezi or powerpoint, etc	5 pts	Used unique features for powerpoint slides. Also followed IEEE format for documentation.
Performed substantial analysis using database techniques Project must include an analytics component	3 pts	Used Tableau to visulaize the data we found that the electricity usage is directly proportional to building area grouped based on the type of building and department.
Used a new database or data warehouse tool not covered in the HW or class	3 pts	Used MongoDB (NOSQL) for querying the data and AWS glue as an ETL tool.
Used appropriate data modeling techniques	5 pts	Used Draw.io and Canva for flowchart and architecture diagram.
Used ETL tool	1 pts	Used AWS glue as a ETL tool.
Demonstrated how Analytics support business decisions	3 pts	Collected data for the electricity usage and we are now able to see their usage trends and understood how much energy was used by various departments.
Used RDBMS Idea is to exercise as many topics from the course as possible	1 pts	We have used RDBMS to run various queries on MySQL Workbench.
Used Datawarehouse Idea is to exercise as many topics from the course as possible	1 pts	Used AWS Redshift for running various querying customers electricity usage in various departments in San Jose.

Includes DB Connectivity / API calls Possibly using Python	1 pts	Using PyMongo library, we connected to the energy consumption database.
Used NOSQL	1 pts	Used MongoDB for analysing the data.