Practicum Final Report (CSE-ISYE-MGT 6748)
Company* Inc

Team 1
Team Member 1, Team Member 2, Yuxi Chen

Georgia Institute of Technology
07/18/2024

1. Introduction

In 2022, Georgia Advanced Technology Ventures began a project to build “Science Square”, next to Georgia Tech’s campus, with an estimation of 10 years until completion [1]. With phase one completed in March 2024, Science Square has four more phases left. Construction projects can range from months to years, depending on the scale of the project. It can be difficult to predict the final costs and time frame. To address this problem, the construction industry refers this task to experts known as “Estimators” to provide this information. Estimators are mainly responsible for the preparation of prices and estimates for building works and primarily the pricing of the BoQ so as to ensure adequate returns on resources employed [3].

To streamline this process, Company* serves as a digital services platform that draws on the expertise and experience of seasoned cost estimating professionals. Company* was founded to address critical shared pain in the professional construction cost estimating industry brought on by a dwindling pool of skilled workers, increased certification demand, and a lack of infrastructure to expedite the acquisition of objective, and unbiased construction cost estimates.

Additionally, Company* also serves as a Vendor Management System (VMS) that connects architects, engineers, government entities, and both public and private institutions with a carefully selected network of experienced cost estimators specializing in various construction areas. Their platform simplifies the construction cost estimating process, ensuring industry-standard quality, reliability, and dependability.

2. Problem Definition

Company* is having difficulty managing deadlines and allocating resources for both full-time employees and contractors. This has resulted in inefficient scheduling for cost estimation tasks and created gaps in service provider workloads. They are seeking to utilize data analytics and machine learning to tackle their project scheduling requirements.

Estimators are provided with design documentation and are given a default of 21 days to complete their estimation. Estimators are paid a flat rate per job, whether they require more or less time than the default number of days. However, if they take less time than required, they can be assigned onto a new project to reduce wait time between projects.

The main objective of this project is to determine the average time of any estimator to complete an estimation based on the number of hours they have reported. Furthermore, to create a visualization tool that would work for both current projects/estimators and future.

Company* is proposing to have the project split into stages from concept, proposal and final stage. Each stage is given approximately three weeks to complete. Each project will be assigned two to ten individuals from the cohort to work on.

3. Literature Review

The construction industry is facing many significant challenges, including a shortage of skilled workers, increasing significant amount of project demands, and unmatched/ delayed support to projects and unstructured cost estimation processes. Company*, the VMS platform is here to solve these painful challenges with their latest technology system. As “The key role of cost estimating in project management” article had stated “Any project, big or small, or regardless of the industry, needs to be performed and delivered under certain constraints. Cost is one of those constraints that project management needs to effectively control. Cost can be the driving force or an impeding factor in determining the future of a project. Therefore, one can easily spot the need for applying thorough cost estimating processes and techniques to ensure that a project can be viable” [5].

On top of the importance, the construction industry as a whole also is ”experiencing a shortage of skilled workers, which has a direct impact on the cost estimating sector. This shortage is attributed to an aging workforce, a decline in vocational training programs, and the perception of construction as an unattractive career choice for younger generations” [2]. Company* would be the perfect way to solve these issues. It offers the latest tools that help to reduce cost and streamline the processes, as well as making client’s projects easier and affordable to match their budgets. Furthermore, Company* also helps to improve efficiency and accuracy in project management,and making clients’ construction projects successful without hesitation.

4. Data overview

4.1. Database access

The datasets were provided through Microsoft Azure Cosmo DB. Azure Cosmos DB is a fully managed NoSQL, relational, and vector database [4]. Initially, we had challenges with accessing the database due to the confidentiality of the data and privacy information. After bypassing the firewall, we were able to gather information and perform exploratory data analysis.

To query the database, we used our individual keys to access Cosmos DB and queried the container using NoSQL, Python and Jupyter notebook. We found that the data were separated into 15 unique partition keys.

```
{'partitionKey': 'account'}
{'partitionKey': 'comment'}
{'partitionKey': 'constructionCategory'}
{'partitionKey': 'content:project'}
{'partitionKey': 'disciplineSkill'}
{'partitionKey': 'email'}
{'partitionKey': 'invitation'}
{'partitionKey': 'leger'}
{'partitionKey': 'proEst:snapshot'}
{'partitionKey': 'project'}
{'partitionKey': 'projectRate'}
{'partitionKey': 'projectType'}
{'partitionKey': 'projectTypeCategory'}
{'partitionKey': 'setting'}
{'partitionKey': 'user'}
```

Figure 1. Partition Keys

4.2. User data

After querying the different keys, we were able to narrow down the data we wanted to use. We found that the datasets from 'project' and 'user', were most relevant to this project. In 'user', it contained all the employees and contractors of Company*. We pulled from the fields of givenName, familyName, id, userType, expertType, primaryDiscipline and expertRole from 'user' data.

givenName	familyName	id	userType	expertType	primaryDiscipline	expertRole
Test	NR	[redacted]				
Nick	[redacted]		0			
Gina				FTE		
Sanjan				FTE		
Sanjan			0			
CeCe			0			
Sanjan			2	FTE	Electrical	CostEstimator

Figure 2: 'user' data

From Figure 2, we can see that there are duplicate names and null values. We decided to keep the duplicated names, as they each had their own unique id. You can find this file as 'cc_names.csv'.

4.3 Project data

In 'project' data, we were able to identify the estimates field. Our initial assumption was to use either effort hours or manual hours as the estimated number of hours used by the estimator. By using project id we could map the individual users to the projects they worked on. By calculating the difference in created date and updated date, we confirmed the three week estimate as stated by Company* for all its estimators. You can find this file as 'project_estimates'.

project_id	phase_type	phase_status	estimates_id	created_date	updated_date	discipline	estimate_status	effort_hours	manual_hours	difficulty
4b6cbb7b-bd7b-	0	2	068dbcf6-36	2022-02-06T0	2022-02-06T0	0	1	1	1	0
4b6cbb7b-bd7b-	0	2	0fba33fa-d81	2022-02-06T0	2022-02-06T0	1	1	8	8	0
4b6cbb7b-bd7b-	0	2	123a6559-ae	2022-02-06T0	2022-02-06T0	2	1	19	19	0
4b6cbb7b-bd7b-	0	2	7b12dc75-33	2022-02-06T0	2022-02-06T0	3	1	19	19	0
4b6cbb7b-bd7b-	0	2	c79a9c17-0c	2022-02-06T0	2022-02-06T0	4	1	19	19	0
4b6cbb7b-bd7b-	0	2	583ea6d8-30	2022-02-06T0	2022-02-06T0	5	1	19	19	0

Figure 3: 'project' estimates

In Figure 3, it can be seen that effort hours and manual hours are identical, but we did find discrepancies in hours. Furthermore, in this portion of the data, we were able to identify the various statuses of the projects.

```
/// <summary>
/// Initial state.
/// </summary>
Draft = 0,

/// <summary>
/// Alternate initial state.
/// </summary>
New = 1,

/// <summary>
/// Item has been submitted by the client.
/// </summary>
Submitted = 2,

/// <summary>
/// Item has been accepted by the provider for initial review and fee generation.
/// </summary>
Accepted = 3,

/// <summary>
/// Phase Only, indicated provider has selected a fee type.
/// </summary>
FeeSelected = 4,

/// <summary>
/// Fee for this item is ready for client review.
/// </summary>
FeeReady = 5,

/// <summary>
/// A fee counteroffer has been sent to the provider by the client.
/// </summary>
FeeCounter = 6,

/// <summary>
/// Client has approved the item, waiting for provider to start work.
/// </summary>
Approved = 7,

/// <summary>
/// Provider has started item work.
/// </summary>
InProgress = 8,

/// <summary>
/// Provider has completed item work. This closes the item.
/// </summary>
Complete = 9,

/// <summary>
/// Provider has rejected item work. This closes the item.
/// </summary>
Rejected = 10,

/// <summary>
/// An in-progress item has been canceled. This closes the item.
/// </summary>
Canceled = 11,
```

Figure 4. Status Summary

Figure 4, explains the various status definitions. We found that from status 8 to 9, is the amount of time the provider or estimator takes to complete the estimate.

project_id	labor_status	user_id	hours
44bf1793-4222-4	7	07199bd7-9430-46c8-b11e-bc9b3	8
44bf1793-4222-4	1	07199bd7-9430-46c8-b11e-bc9b3	22
44bf1793-4222-4	7	07199bd7-9430-46c8-b11e-bc9b3	8
0ff662f8-1a9f-4a	1	2525f3e8-b92c-4628-8e81-0d3d3	8
0ff662f8-1a9f-4a	1	2525f3e8-b92c-4628-8e81-0d3d3	8
0ff662f8-1a9f-4a	1	2525f3e8-b92c-4628-8e81-0d3d3	16
0ff662f8-1a9f-4a	1	07199bd7-9430-46c8-b11e-bc9b3	16

Figure 5. 'Project' labor

Company* later provided additional feedback, for us to look into the labor field nested within estimates. From here, we were able to attribute each user to the projects they worked on, the number of hours spent on each project and the current status of the project. This file can be found as 'labor_hours'. We then merged 'labor_hours' with 'cc_names' to create 'user_hours'.

5. Experimental Observations

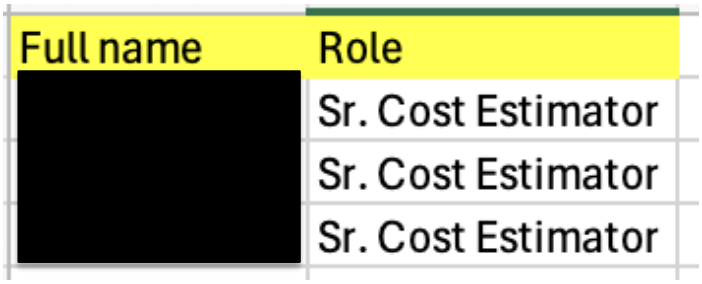
In our analysis of the cleaned data, we made several reasonable assumptions in order to ensure the mapping process is accurate and precise. Some of the estimators/employees are adjusted based on reasonable modification to their designated roles. In the beginning, our dataset revealed that 20 individuals out of 26 directly matched the specified positions. This prompted us to dive deeper into our raw data and Azure database to ensure accuracy and completeness in our role assignments. Figure 6 shows the mapping we have done for both labor and role databases.

project_id	labor_status	user_id	hours	givenName	familyName	Full name	Role
44bf1793-4222-4790-92e0-67ae0a4b6a8d	7	07199bd7-9430-46c8-b11e-bc9b3737aa83	8	Sanjan	[redacted]		Electrical
44bf1793-4222-4790-92e0-67ae0a4b6a8d	1	07199bd7-9430-46c8-b11e-bc9b3737aa83	22	Sanjan			Electrical
44bf1793-4222-4790-92e0-67ae0a4b6a8d	7	07199bd7-9430-46c8-b11e-bc9b3737aa83	8	Sanjan			Electrical

Figure 6. Mapping for Labor and Roles

After further investigation, we discovered that out of six unmatchable employees, some were listed under different names, which initially caused errors in our mapping. For example, Sanjan Adhikari, appeared in one part of the data, yet, in the provided role list dataset, he was listed as Sanjan Estimator. Furthermore, , well-known as Employee 1* in our Slack group, holds the Principal role. On top of these two, Employee 2*, whose responsibilities include overseeing all projects, should rightfully be classified under the Principal role due to the scope of her duties and authority.

For the other three unmatchable estimators based on their full name, we used other cleaned data sources to map them with the reasonable roles. The project ID, their job stage/status and hours that worked all aligned with the job function for Sr. Estimator. Figure 7 shows the three unmatchable estimators in detail.



Full name	Role
[redacted]	Sr. Cost Estimator
	Sr. Cost Estimator
	Sr. Cost Estimator

Figure 7. Unmatchable Estimators

Lastly, we also saw a common placeholder/name from the list appear in the cleaned data user_hours_Final Data.xlsx and CC_Export Time-Project.xlsx from the Company* database as well. It is typical for companies to create testing placeholders when mapping and testing new systems. These placeholders are necessary for companies to validate the accuracy of their new system, however, usually, companies will consider these testing placeholders as the outliers for data analysis to maintain precise and accurate results. In our project with Company*, the entry 'Test NR' would serve as a testing placeholder. Recognizing 'Test NR' as an outlier, we excluded it from our core analysis to maintain the integrity and accuracy of our project.

On top of Test NR, Figure 8 also indicates some projects that do not have assigned estimators, this might be potentially caused by the labor status, where either it was canceled or still in the initial stage from clients' side or system errors. Based on the validation process, we decided to count these as outliers.

project_id	labor_status	user_id	hours	givenName	familyName	Full name	Role
c69d2a5f-62f5-48cf-94da-a5d122738191	11		0				outliers
24836fc7-6c5e-4a09-bf9b-b0a472b0bbf4	1		0				outliers
71b594b6-64a8-4c11-aa61-a1c95c5f9f60	1		0				outliers
71b594b6-64a8-4c11-aa61-a1c95c5f9f60	1		0				outliers
efdfdf08-98fb-4e4f-b019-53737ea1be7c	1		0				outliers

Figure 8. Projects without Assigned Estimators Classified as Outliers

By investigating and adjusting for these datasets, we now have a more accurate and representative mapping of roles, which is crucial for further analysis and findings. This process indicates the importance of data verification, a deeper dive into data scrubbing, and the consideration of potential mistakes and outliers in names and placeholders in large datasets.

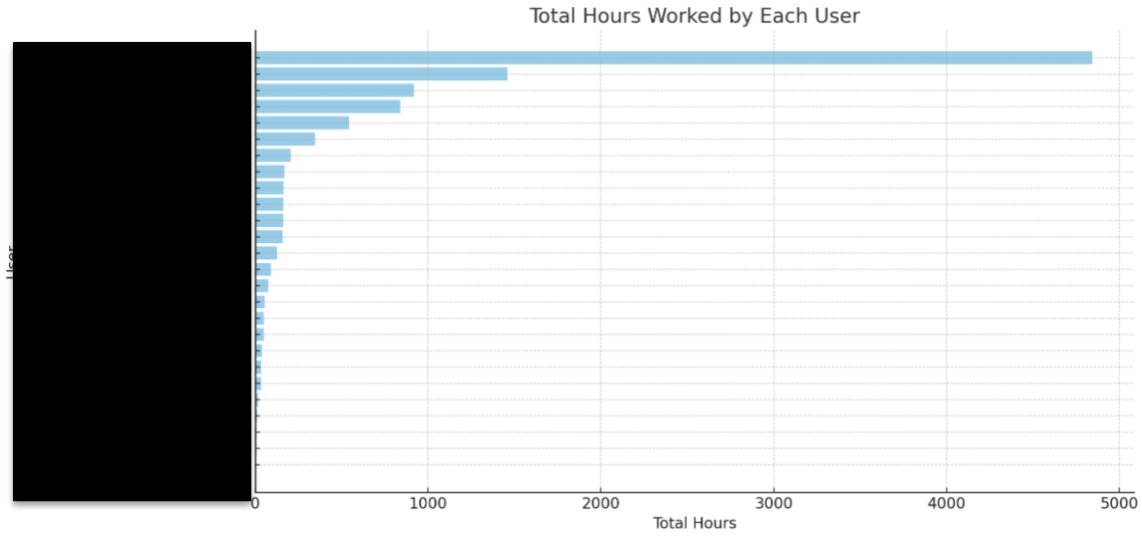


Figure 9. Total Hours by Each Estimators

The chart above Figure 9 indicates the total hours worked by each user according to our database. We can see that as the given database, Employee 1*has worked the most hours, and from Employee 4*, we can tell that the total hours tend to be very steady and gradually decrease among all estimators. Lee's total hours could be affected by the total workload/projects, Figure 10 below would help with the breakdown.

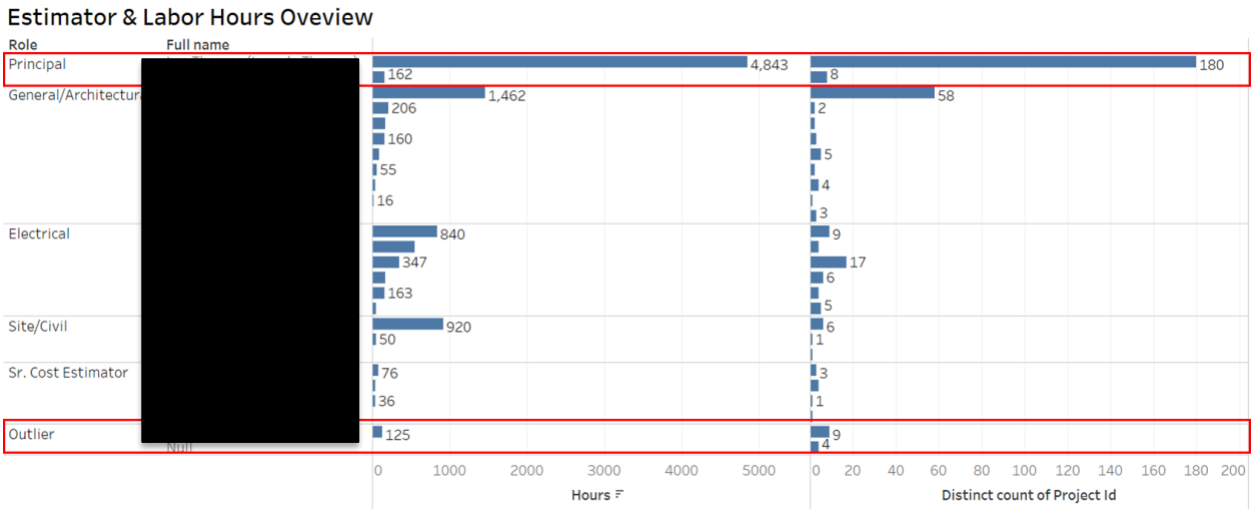


Figure 10. Breakdown of Roles and Estimators by Hours and Project ID

Plotted each estimator with respect to their roles, Hours and count distinct from Project ID. From the chart above, it is observed that Employee 1* and Employee 2* are in Principal roles, with Employee 1* appearing in most projects. Also observed there are hours with Test NR and Null as estimator names. These are what we excluded from the next visualizations as they skew the data and do not hold the roles that we are going to analyze.

After removing the outliers, Figure 11, 12, 13 below is the overview breakdown of number of hours, estimators and project based on the roles. It can be seen that for the role General/Architectural has the highest number of estimators (9 people), labor hours (2,201 hours) and count of projects (72 projects). Followed by Electrical, Sr Cost Estimator and Site/Civil.

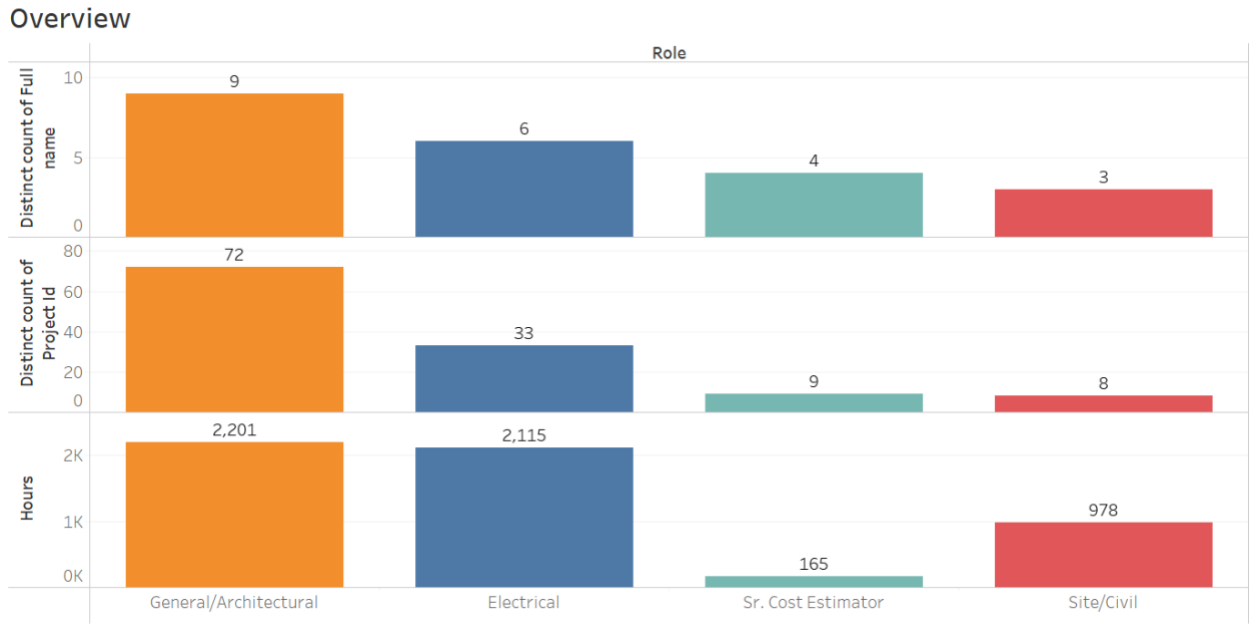


Figure 11. Overview of Roles in Hours, Project ID and Estimator

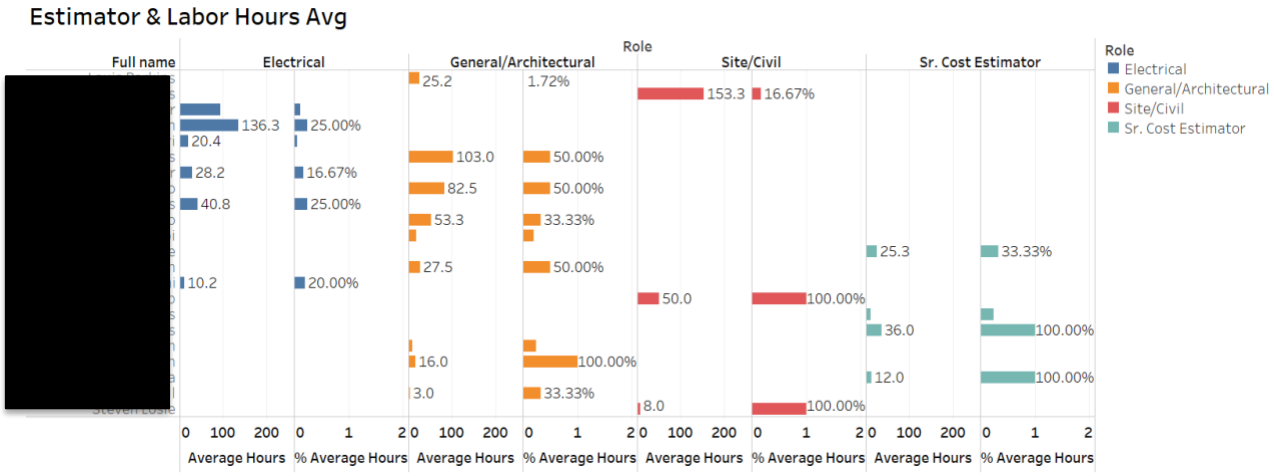


Figure 12. Average Hours and % Average Hours for Each Estimator and Roles

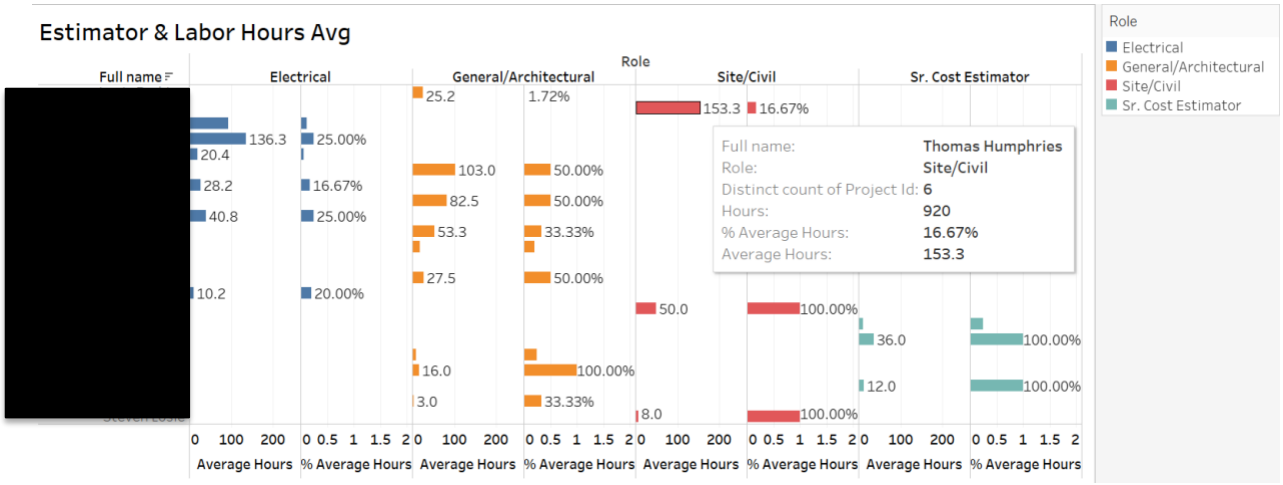


Figure 13. Tooltip Details of The Calculation for Average Hours

Visualization above shows the breakdown of all the roles and estimators by average hours and % average hours they put in after excluding the outliers. Average Hours is defined by total hours by the estimator divided by count distinct Project ID they are involved in. In the example above, Employee 3* contributed a total of 920 hours for 6 Project IDs. His average hours is 920 divided by 6, 153.3 hours.

%Average Hours is calculated by average hours divided by sum of total hours. In this example, Employee 3* has 153.3 hours divided by 920 hours. This gives 16.67%. It means, for each project, on average he contributed 16.67% of his total hours.

Table of Estimator & Labor Hours Avg

	Role			
	Electrical	General/Architectural	Site/Civil	Sr. Cost Estimator
Hours	2,115	2,201	978	165
Distinct count of Project Id	33	72	8	9
Average Hours Per Project	64.09	30.56	122.25	18.33
% Average Per Project	3.03%	1.39%	12.50%	11.11%
Average Hrs/Role/Estimator/Project	10.68	3.40	40.75	4.58
Distinct count of Full name	6	9	3	4

Table 1. Table of Estimator & Labor Hours Avg

The Table 1 details further the hours spent per role per estimator per project. Provided the definition in table below

Statistics	Definition
Hours	Total hours spent for respective roles
Distinct count of Project ID	Unique count of Project ID
Average Hours Per Project	Hours / Distinct count of Project ID

% Average Per Project	Average Hours Per Project / Hours * 100%
Average Hrs/Role/Estimator/Project	Hours / Distinct count of Project ID / Distinct count of Full Name
Distinct count of Full Name	Unique count of estimator

On average, each person will spend about 10.68 hours as Electrical, 3.4 hours as General/Architectural, 40.75 hours as Site/Civil and 4.58 hours as Sr. Cost Estimator per project.

6. Conclusion

6.1 Discussion

From the overview, it is observed that the General/Architectural role has the most estimators (with a total of 9 estimators), the highest labor hours, and the count of projects. However, after deep diving into the data visualization, the role Site/Civil has consumed the most time spent per estimator at 40.75 hours. Site/Civil only has three estimators. In other words, each site/civil estimator could only work on one project at the most per week. Noting, this is the average, so actual working hours for Site/Civil will vary upon the level of difficulty of the project. Furthermore, based on the tableau dashboard above, we can conclude that Site/Civil made up the majority of working hours in a Company* project. This information suggests that there is a lack of manpower in this role.

6.2 Difficulties

In this project, we encountered a few challenges. The first challenge was getting every member of our team access to the data, due to the Cosmos DB firewall and privacy settings. We then had difficulty with how the terminology was used in relation to the scope of the project. For instance, ‘cost engineer’ versus ‘estimator’. Lastly, we had trouble with identifying the correct fields to find the hours inputted by the estimator. We identified various fields that could be correlated to the number of hours performed and also noted that there were estimators with null values in their labor hours. Due to the timeframe and limited dataset, of the project, we were not able to create a predictive model.

6.3 Next Steps

In the future continuation of the project, we would suggest future works to include

6.3.1 Predictive Analytics

Train machine learning models to predict future estimator needs and project timelines. Use regression models, time series analysis, or classification algorithms as appropriate.

6.3.2 Prescriptive Analytics

Develop optimization models to recommend optimal resource allocation and project scheduling. Use linear programming or other optimization techniques to solve for the best allocation of resources.

6.3.3 Real-time Analytics

Set up real-time data feeds to monitor ongoing project activities. Create dashboards to visualize real-time data and provide actionable insights. This is achievable once we get the connection to a real-time database into the visualization tool.

7. Work Distribution

MSA Practicum - Summary of Workload Distribution

Task	Description	Team Member Contributions
EDA	Understanding the datasets and gaining insights into the structure of the data.	Fiona (Fangfei) Li
Data Cleaning / Transformation	Cleaning datasets to meet the assumptions of statistical models	Yuxi Chen
Methodology	Examining and understanding data to uncover underlying patterns, relationships, and insights	Jessy
Analysis and Results	Analyzing the insights that we gain from methodology and preparing the data for our final report/ project	All (Jessy, Fiona (Fangfei) Li, Yuxi Chen)
Midterm Report	A ppt report showing what we have done, what needs to be done, and project overview	All (Jessy, Fiona (Fangfei) Li, Yuxi Chen)
Final Report	A report showing what we have achieved, and analysis based on the project	All (Jessy, Fiona (Fangfei) Li, Yuxi Chen)
Project Management	To arrange on group meeting, track deadlines and submit to Canvas	Jessy

8. References

1. Green, J. (2022, September 16). *A decade in the making, Science Square Project Has Official Start Date*. Urbanize Atlanta.

https://atlanta.urbanize.city/post/science-square-georgia-tech-development-project-has-official-start-date
2. Miller, R. (2020). Vocational training and the future of the construction workforce. *Labor Market Studies*, 58(1), 88-104
3. Rajpatty, S. J. (2008). The Role of the Estimator in Today's Construction Industry. *AACE International Transactions*, DE121.
4. seesharprun. (n.d.). *Introduction to Azure Cosmos DB*. Learn.microsoft.com. https://learn.microsoft.com/en-us/azure/cosmos-db/introduction
5. *The key role of Cost Estimating in Project Management*. Cost Estimating & Project Controls. (26AD). https://www.costengineering.eu/blog-article/the-key-role-of-cost-estimating-in-project-management