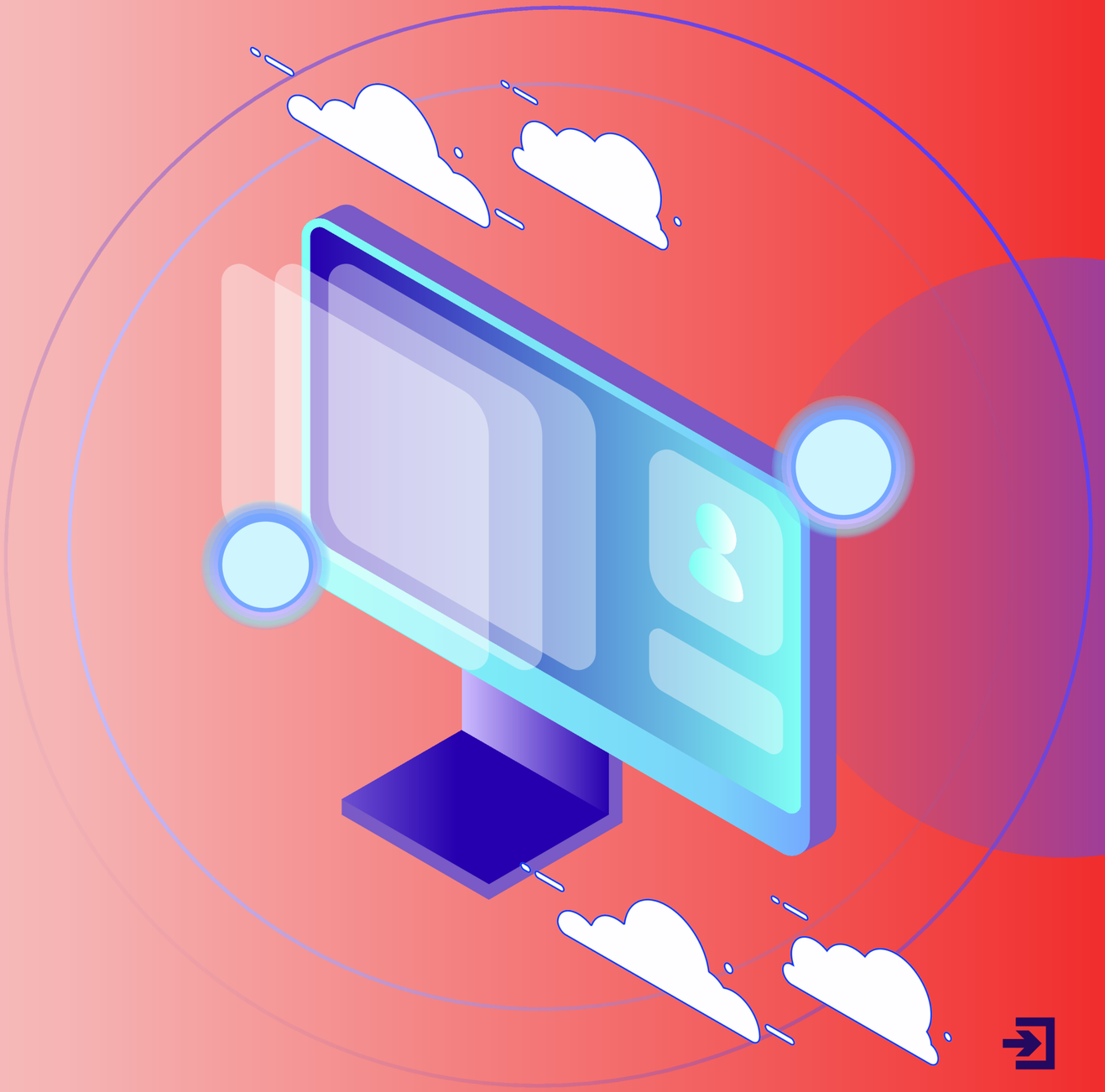
ML COMPETITION

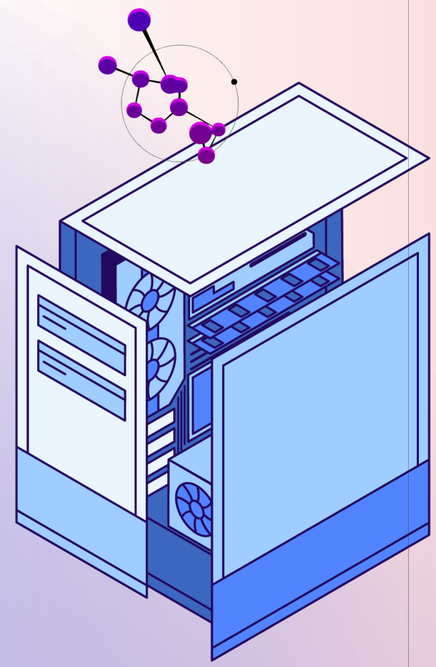
Michelle Cheung | Carlos Sujanto

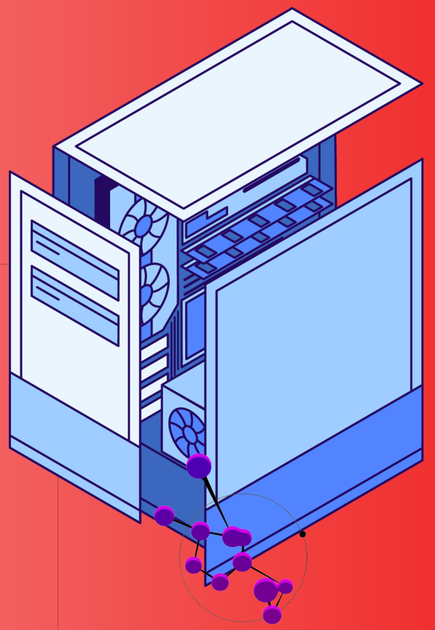


URBAN

SYSTEMS

PROBLEM STATEMENT

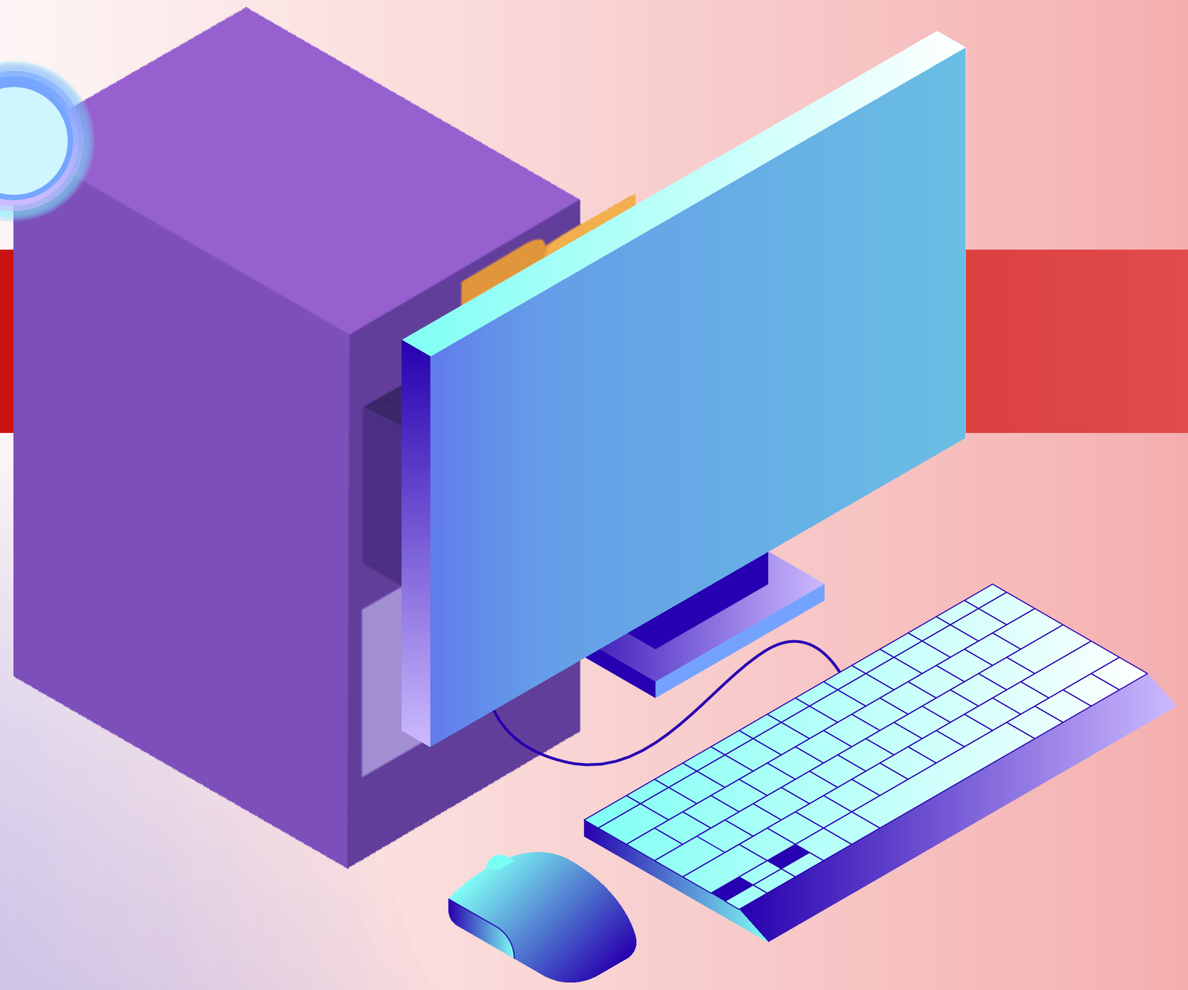




We believe that

predicting revenue generated from projects

would be an invaluable tool in helping understand revenue drivers to enhance decision-making.



TECH STACK

We used a combination of tools. For the database, we used **Neo4J**.

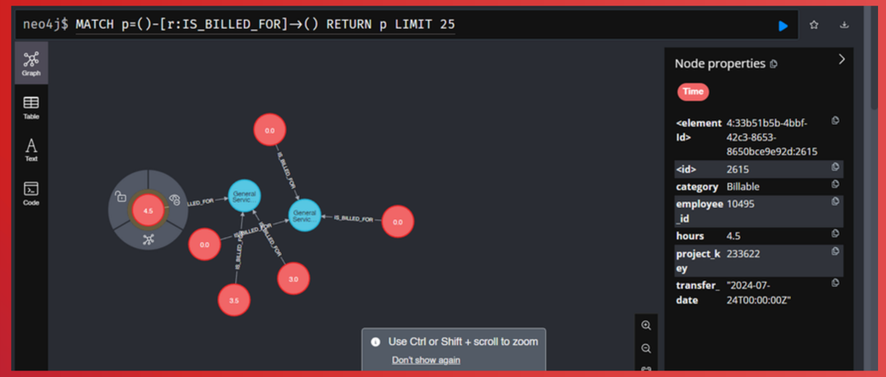
Our script we used **Python**.

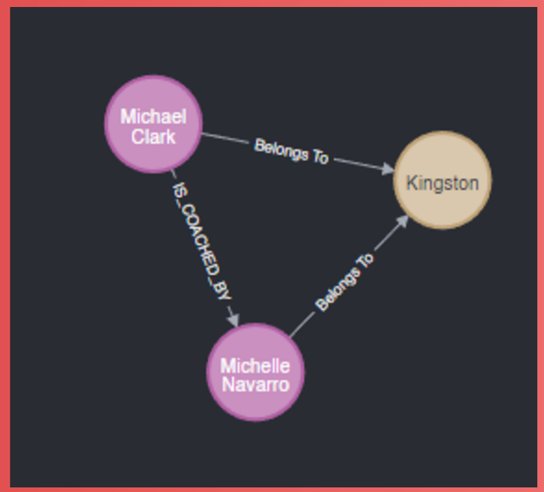
For visualization of data exploration to draw meaningful data:

- **Seaborn**
- **Matplotlib**
- **Pandas**

For Machine Learning:

- **Scikit-learn**


neo4j$ MATCH p=()-[r:IS_BILLED_FOR]→() RETURN p LIMIT 25
Graph
Table
Text
Code
Node properties
Time
<element id> 4:33b51b5b-4bbf-42c3-8653-8650bce9e92d:2615
<id> 2615
category Billable
employee_id 10495
hours 4.5
project_key 233622
transfer_date "2024-07-24T00:00:00Z"
Use Ctrl or Shift + scroll to zoom
Don't show again


Michael Clark
Belongs To
Kingston
IS_COACHED_BY
Belongs To
Michelle Navarro

METHODOLOGY STEPS

Database - Neo4J:

- Read CSV file and created nodes
- Connected each node to another by named relationships

Data Exploration:

- The time spent on the each project type
- The total revenue generated for each project type
- Outliars or trends relating to the revenue of projects

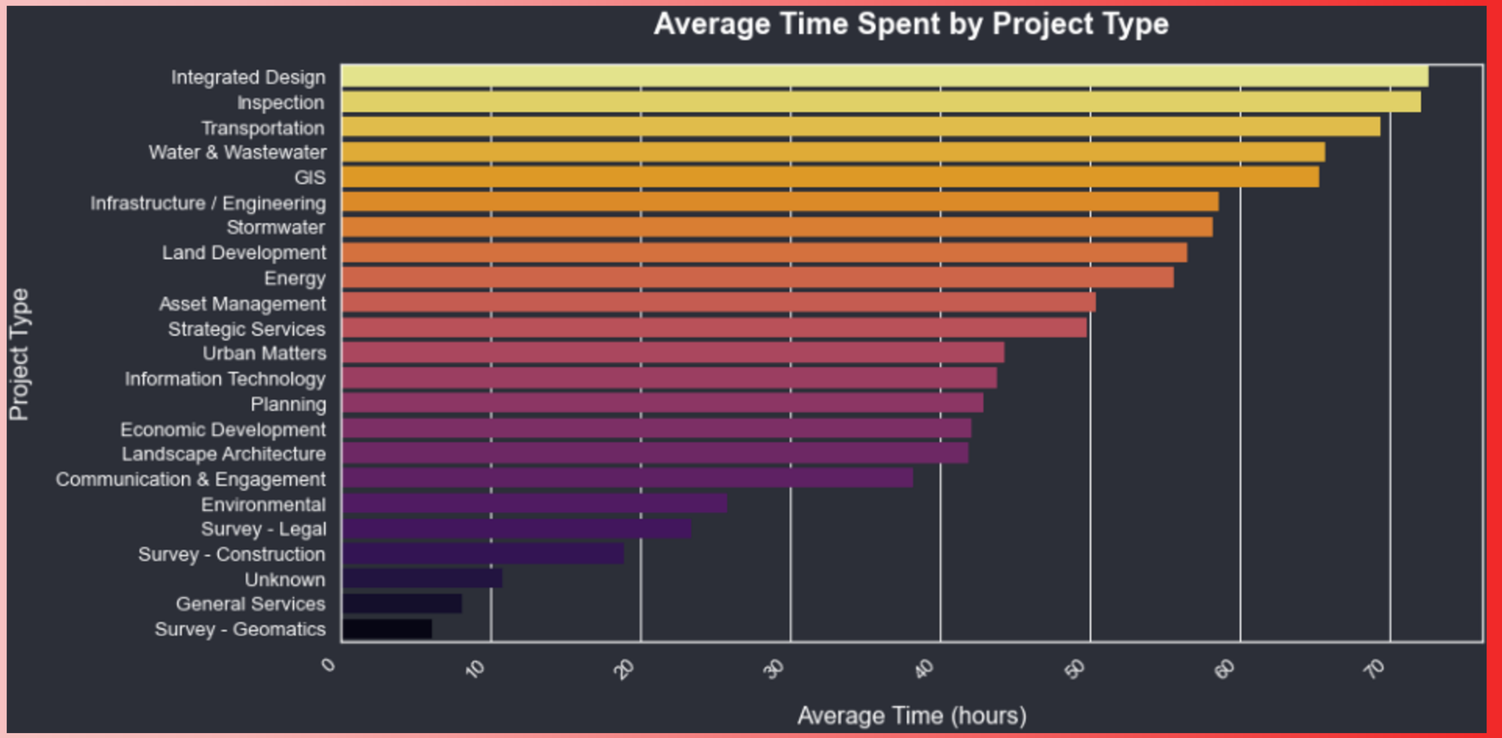
Machine Learning:

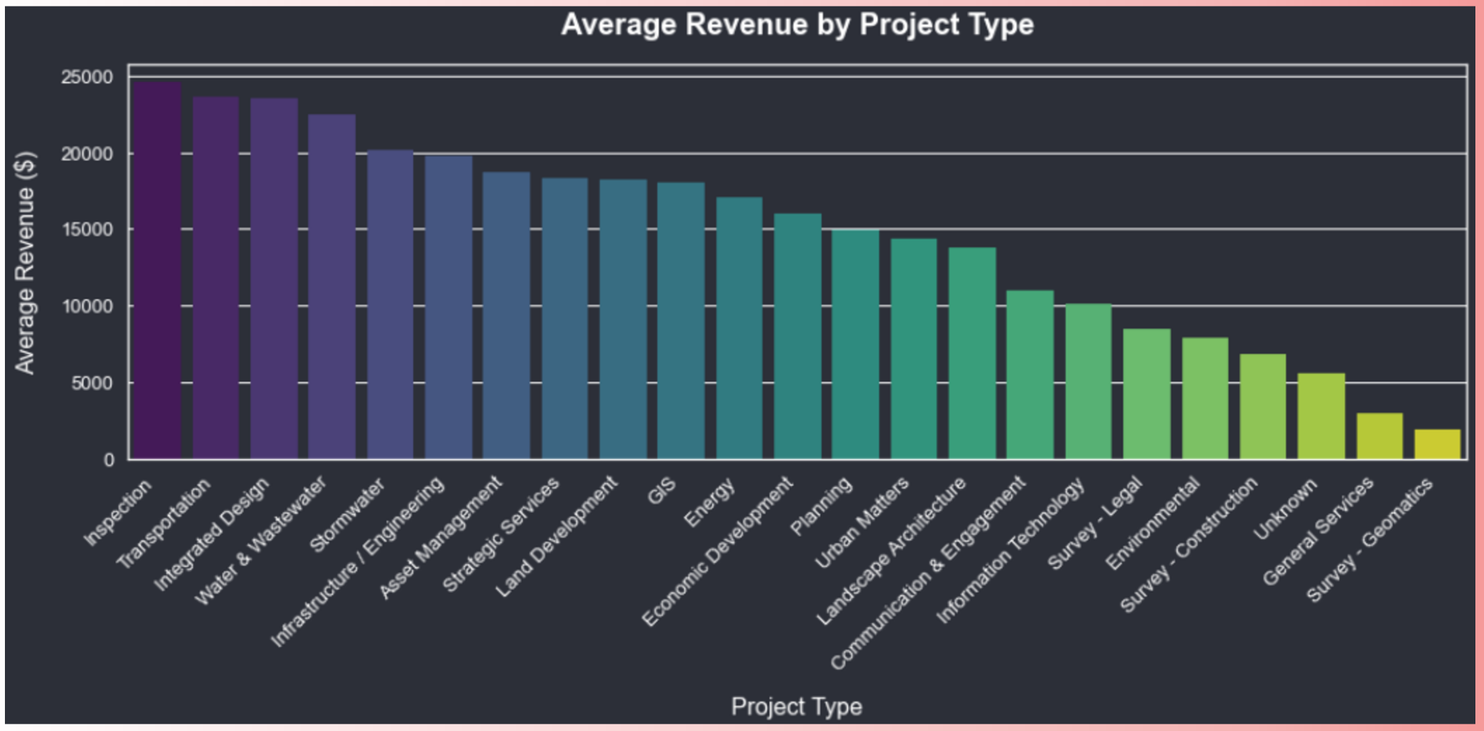
- Utilized a **Random Forest Regressor** to predict the revenue a potential project would be generate
  - Performed **HyperParameter Tuning** to find the best parameters
  - **One-hot encodings** to feed into the Random Forest Regressor to encode all categorical labels (project types)

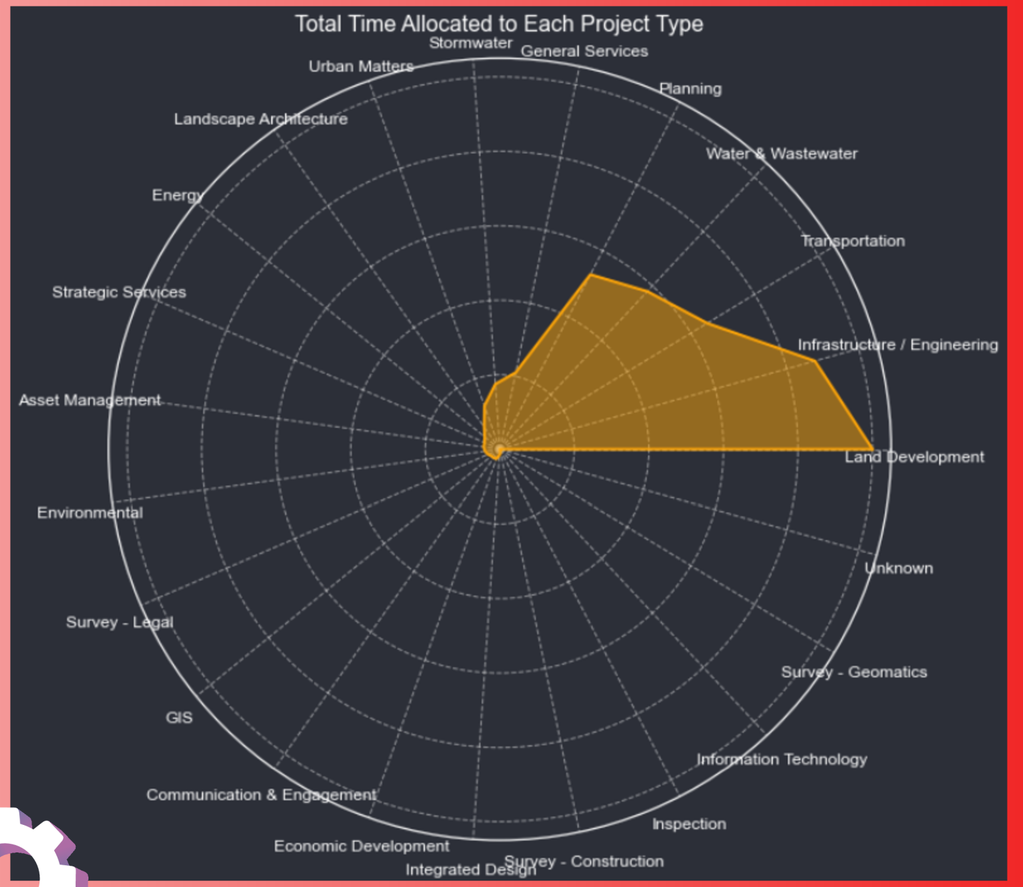
EXPLORATION FINDINGS

TIME SPENT

We found that the **Integrated Design** project type often used the **most** amount of time where as **Survey - Geomatics** used the **least** amount of time. Naturally, this prompted us to look into the revenue generated for these projects.





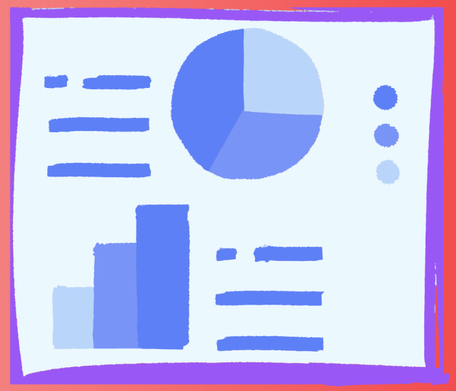


EXPLORATION FINDINGS

REVENUE

Looking at the average costs, we **Survey - Geomatics** had the **lowest revenue**, whereas the **Inspection** project types generated the **highest** average revenues, which contrasts the amount of time spent.

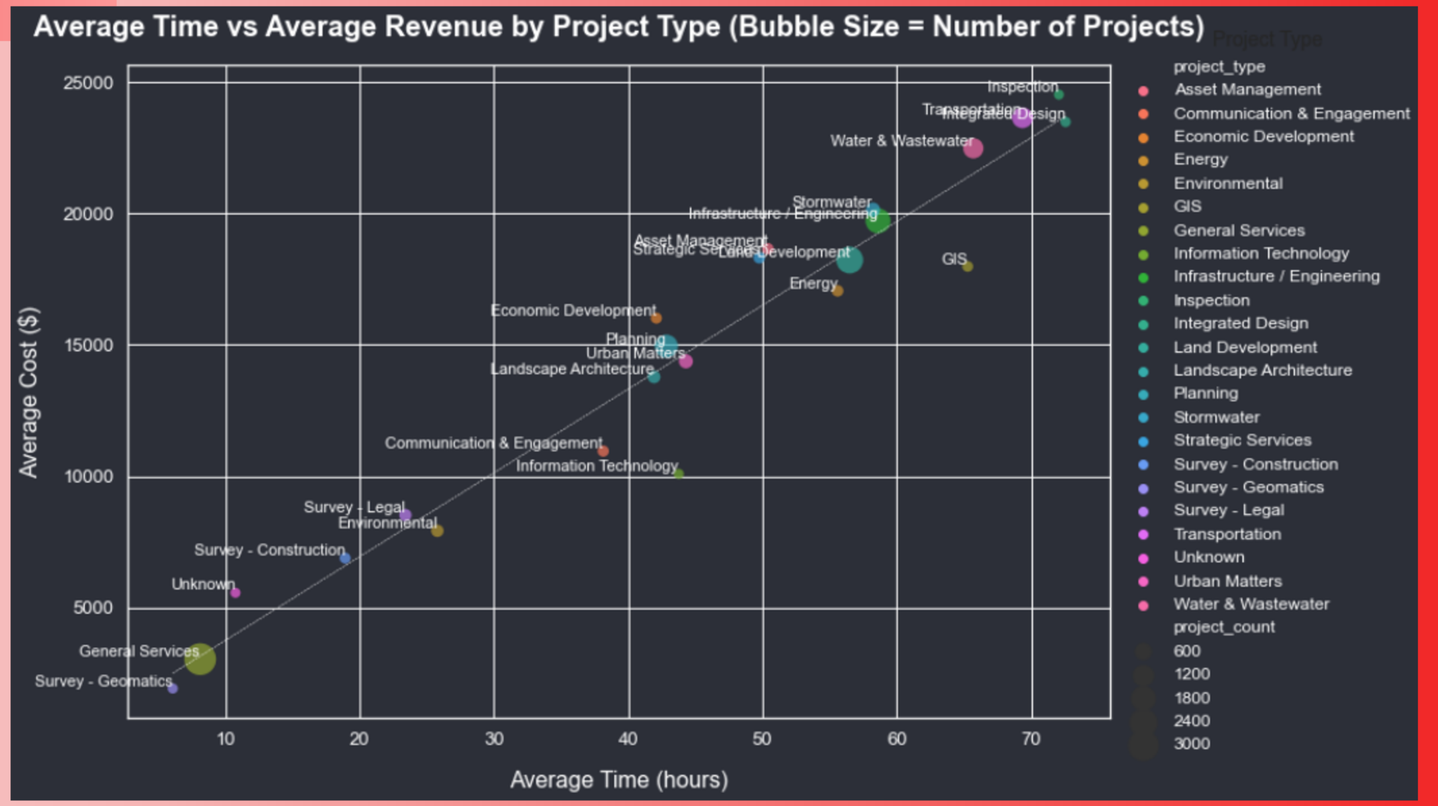
EXPLORATION FINDINGS



TIME VS REVENUE

For the most part, and expected, as the time increases, the generated revenue increases. The only straight-away outlier we see if **GIS** which has a lower revenue even with higher hours.

Additionally, we see that **General Services** has a large amount of projects, but with low time and revenue, whereas Inspection projects have few projects but have huge cost and time commitments.



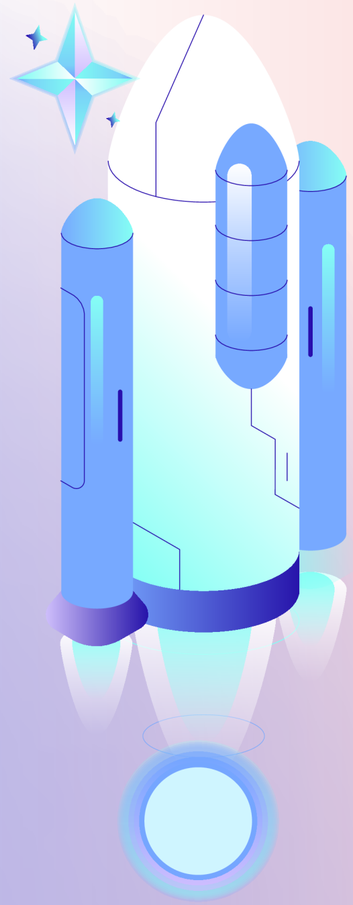
ML RESULTS

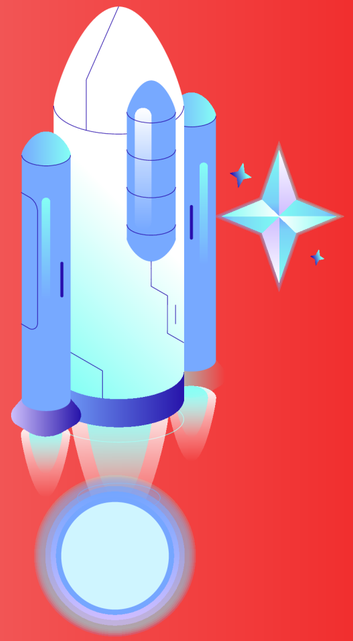
EVALUATION METRICS

Mean Squared Error: 108,813,662.17
R-squared: 0.92
Mean Absolute Error: 2603.85

ML CHALLENGES

When we initially ran our model, our model was severely overfitted with a R-Squared value of 0.99. We attempted to perform various methods of checking for overfitting including predicting on trained data, 5-fold cross validation. In the end, changing our parameters helped us achieve a R-Squared value of 0.92.





FINAL THOUGHTS

Being completely new to Neo4J and Machine Learning, this was a great opportunity to have hands-on experience working with Data Science tools. We both learned a lot from this competition and would like to thank Urban Systems and DSMLC for hosting the competition.

THANK YOU!

Michelle Cheung | Carlos Sujanto

https://www.linkedin.com/in/michellewycheung/
https://www.linkedin.com/in/carlos-sujanto/

wycmichelle11@hotmail.com
carlos.sujanto@ucalgary.ca