CSCI E-109B: Project Milestone #2

EDA

Group #48

April 13, 2020

Team Members **Topic E - Predicting Project Success**

Michael Lee	**Micah Nickerson**	**Daniel Olal**
mil726@g.harvard.edu	min021@g.harvard.edu	dolal@crystal.harvard.edu

Project TF

Brandon Walker

Project Topic

Successful project planning is a critical part of the success of any business or organization in order to minimize unanticipated delays or cost overruns. With our data science models, we strive to predict project success or failure based on historical project planning records.

For this analysis, we define project success as a combination of two metrics: **budgetary success** and **schedule success**. Budgetary success occurs when the project is completed **on budget***, and schedule success occurs when the project is completed **on schedule***.

Description of Data Set

Our dataset is a collection of capital projects provided by the Mayor's Office of Operations in New York City. The projects in the database include projects managed by NYC city agencies which have a budget of $25 million or more. The information in the dataset includes the date of the project status update, the project ID (PID), the project name, the project description, and the category of the project. Details about the borough, managing agency, client agency are also provided. The progress of the project is defined based on variables including the current phase of the project, design start date, budget forecast, the latest budget changes, the total budget changes, forecast completion, latest schedule changes, and total schedule changes. The dataset contains *multiple rows per project* (per PID), providing time stamped

*- on budget and on schedule indicates that the project is completed within an interval of acceptable deviation, with upper and lower bounds still to be determined.

sequential updates to each of the forecast completion, latest budget changes and latest schedule changes.

A summary of project performance within the dataset is listed in Table 1 below.

Table 1 - Data Distribution (by Project Success Measurement)

Subset of Data	Total (Percentage of Dataset)
Total Projects	**378 (100.00%)**
Projects *Over Budget*	264 *(69.84%)*
Project *Over Schedule*	267 *(70.63%)*
Projects ***Both*** *Over Budget and Over Schedule*	233 *(61.64%)*

Exploratory Data Analysis

Key Questions:

1. ***Given everything you have learned, if you faced this data set in the wild, how would you proceed?***

   The first step of our exploratory data analysis would be to understand the structure of the dataset. We would learn the background of the dataset and how each variable in the dataset is defined. We would then perform **data cleaning** and **standardization** to ensure that the variables were scaled to a mean of zero and a standard deviation of 1. This is important because of the wide range of budget forecast values; some projects have a budget in the $10 million order-of-magnitude while others are in the $1 billion order-of-magnitude. Once the data is cleaned, we would then plot histograms and perform clustering to find the underlying structure of patterns. We would also tag the dataset by binary flags and descriptor tags that enumerate and classify the data based on schedule and cost information. The results of these steps are below.

2. ***What are the important measures?***

   In our review of the assigned papers corresponding to this dataset, we learned that **Earned Value** is a commonly used metric to measure project success (Lukas 2012). Earned Value measures the Schedule Performance Index (SPI) and Cost Performance Index (CPI), thus a project with a higher SPI and CPI would have a higher Earned Value. However, these metrics depend on granular information on a monthly or periodic basis about the actual value of material installed on a construction project and how much money was spent to install that material. **In this dataset, we don't have enough information to calculate the SPI or CPI.** Without the Schedule of Values (SOV) for each project in the database, we can't measure the Earned Value with any level of accuracy so we opted not to use Earned Value for this project. Instead, we propose to use the **absolute percentage error** for **budgetary success** and **schedule success** as our metrics of project success because these values can be calculated from the dataset.

We also considered word frequency and text analysis as an important measure, but based on the provided reading materials, we found that word frequency didn't appear to explain any of the variance for budget or schedule success (example: the word 'park' being most frequent for *both* successful and unsuccessful projects) . Therefore, we decided not to include word frequency as an important metric to evaluate in this analysis. We will consider applying Natural Language Processing and proceed with caution using project descriptions to predict schedule success.

3. ***What are the right questions to ask, and how can the data answer them?***

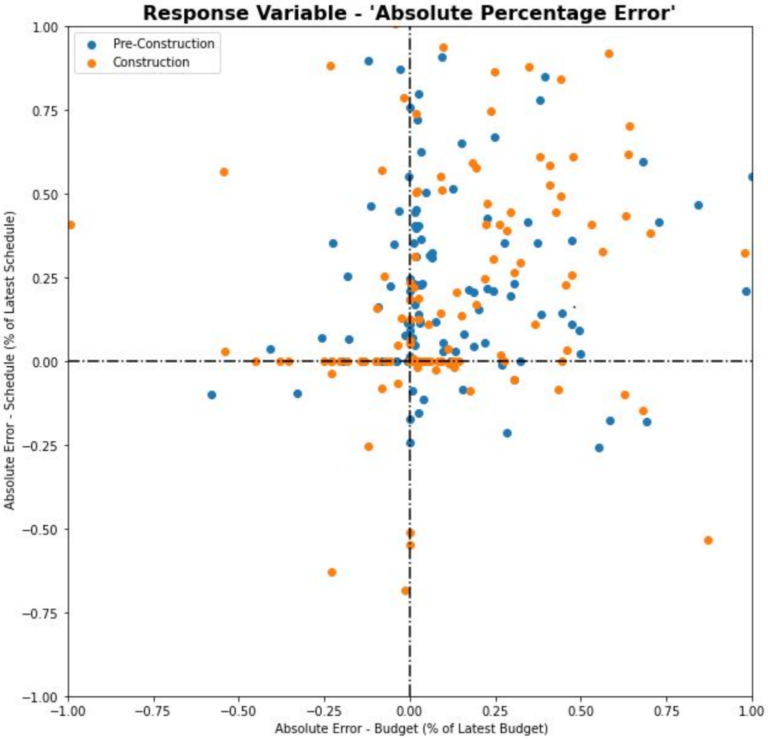
Based on our review of the dataset, we had several questions that we would like to investigate:

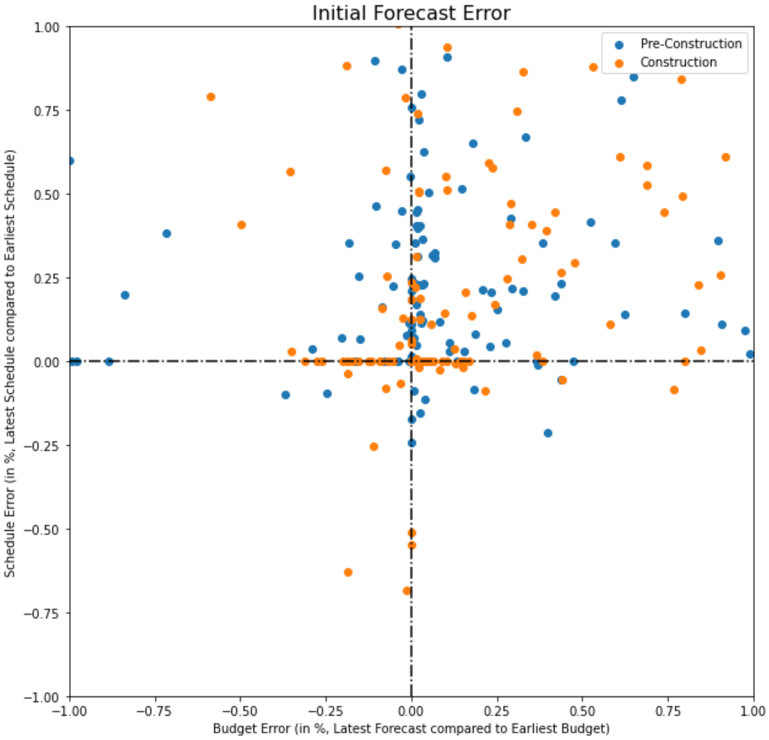
- **Is** project success based on a time series analysis? We can evaluate this question by:
    - performing an ARIMA model.
    - verifying our models with auto-correlation plots (ACF/PACF).
- **Will** LSTMs or GRUs be more effective in helping us understand the data compared to additive models such as GAMs? Do RNN models need to be bidirectional to accommodate for different elapsed phase lengths and start dates?
- **Can** we adequately predict the level of which a project will be over budget or over schedule? With sufficient data, we could build a variational autoencoder or reinforcement learning model to predict what a successful project might look like.
- **Does** geographical location play a role in the success of a project? We can evaluate this in our dataset by separating the data by city borough.
- **Do** certain city agencies manage projects better than others? We can evaluate this in our dataset by separating the data by managing agency and evaluating differences in project success.
- **Is** the rate of project success (based on budgetary and schedule success) dependent on the type of capital project? We can evaluate our dataset to see if certain types of projects, such as transportation, have different levels of project success compared to other types of capital projects such as parks or schools.

Response Variable

Our response variables for measuring project success are budgetary success and schedule success. We propose to measure budgetary success and schedule success as the **absolute percentage error**. For *project budgetary success*, the absolute percentage error is defined as the absolute value of the

total budget changes divided by the most recent project **budget forecast**. For *project schedule success*, the absolute percentage error is defined as the absolute value of the **total schedule changes** divided by the most recent **forecast completion** date.





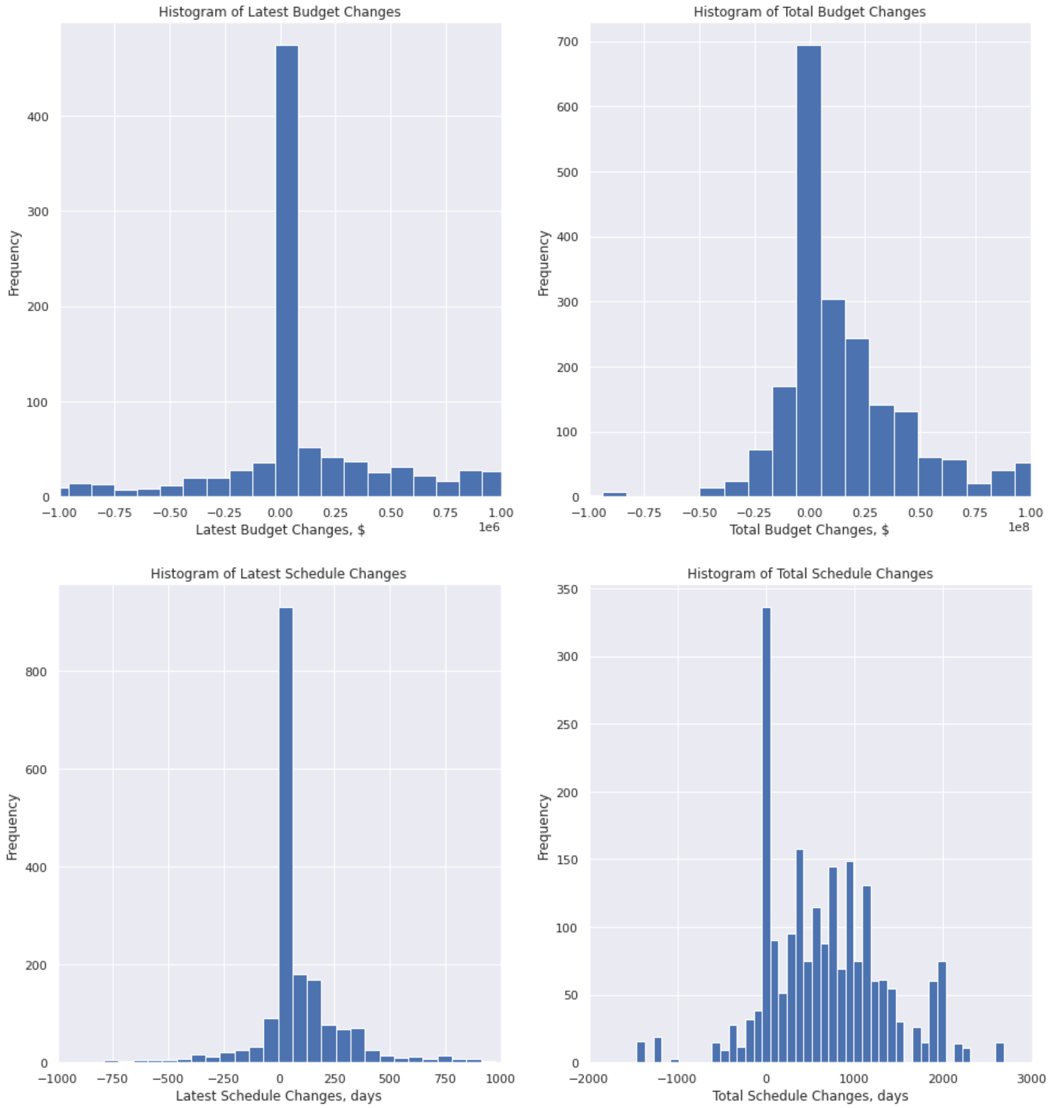
- Above we **plotted the response variable** to understand the distribution of that response. To better understand how forecast date affects the response, above is a comparison of error calculations, using 'latest' versus 'original' construction phase forecasts. There are slight differences between them, but the metrics perform similarly - indicating **using the latest forecast to calculate the response variable is satisfactory**.
- Initial review of plots imply that **pre-construction projects are more likely to have variance in schedule**, while **projects in construction are more likely to have variance in budget** (the congregation of dots on the 0,0 center cross). Further plots by time may indicate this is due to project age.

Data Cleaning

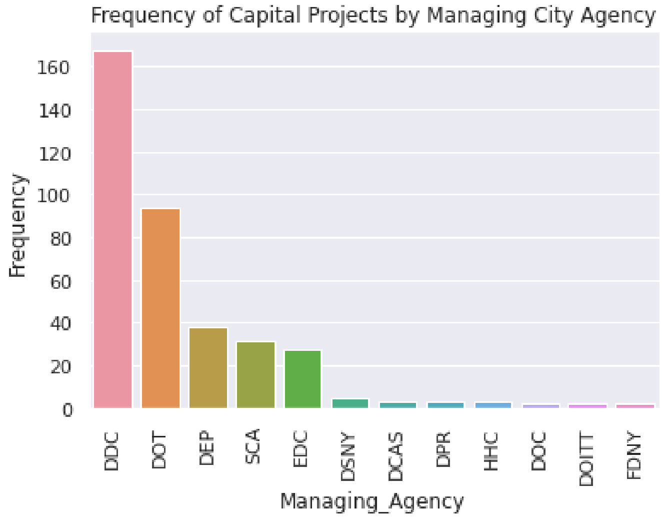
- **What did we do?**
  - We **formatted** and cleaned by dropping missingness from total project budget and schedule (incremental updates are ok), and grouped the dataset into a list of unique project IDs. We **elaborated** the dataset with labels from NYC Open Data, and additional time series deltas (count of days) based on key project milestones. We **classified** the set by adding binary flags for performance / construction status.

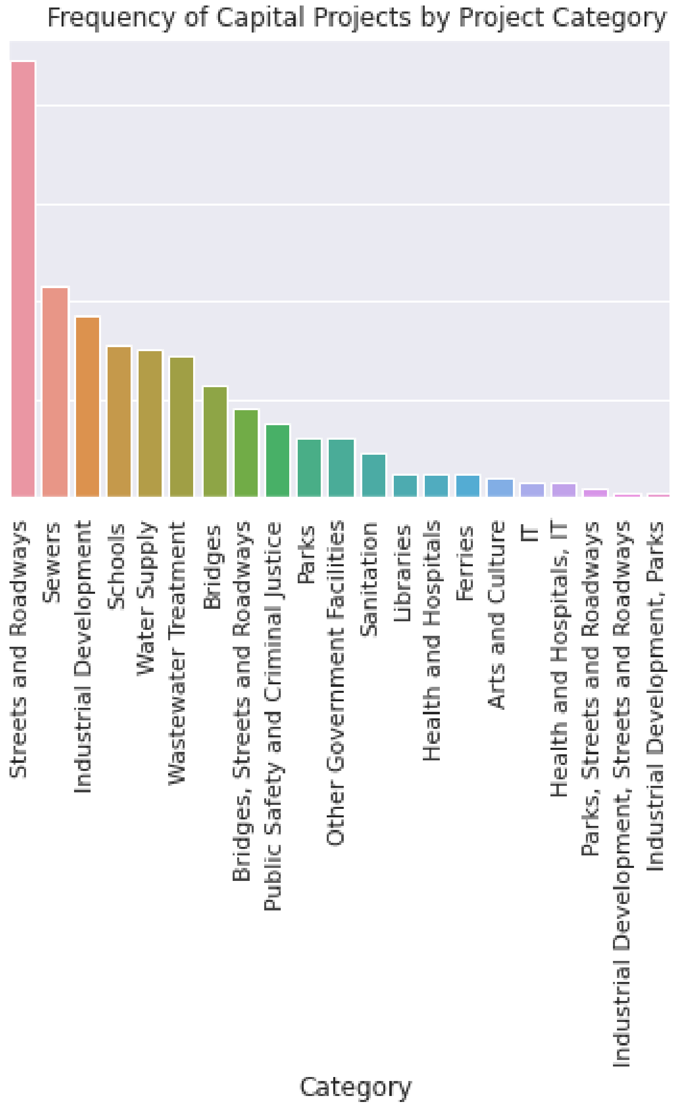
- **Future actions**
    - **Data balancing** - more than half of the dataset contains projects that are both over budget and over schedule. Data augmentation including **upsampling** to balance the number of "successful" projects will be required. There do not appear to be enough projects that satisfactorily meet the absolute percentage error to properly train any data science method (GAM, RNN) in "what is a successful project". Likewise, data balancing is needed to balance geographic locations (downsample Brooklyn and/or upsample Queens) to be about the same.
    - **Data normalization** - since projects have vastly different statuses (some are in construction, others are only in planning), start dates and elapsed time since first recorded projection (time or budget), we have to normalize each for **elapsed time** "time = 0" in order to compare them. Due to the lack of data from "start_of_design" till "first_schedule_update", models may also require **padding**.

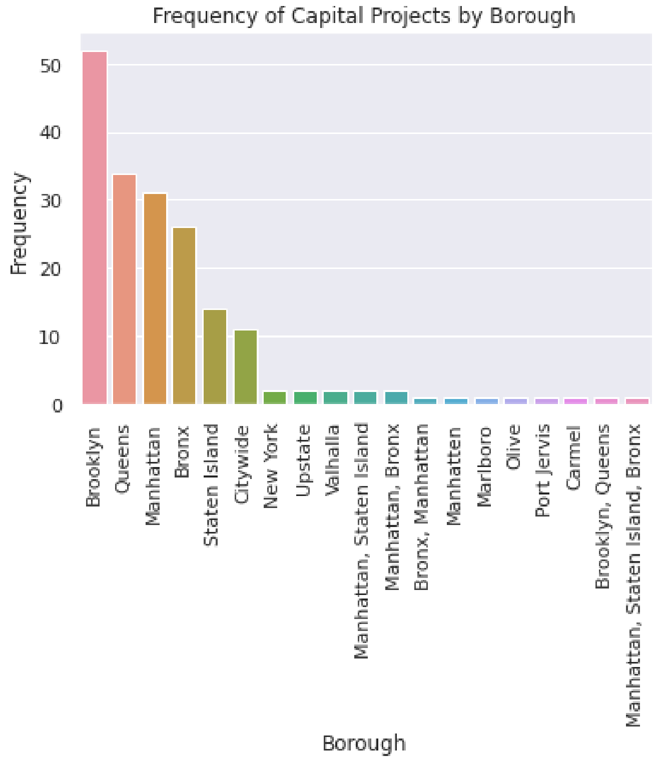
Data Exploration



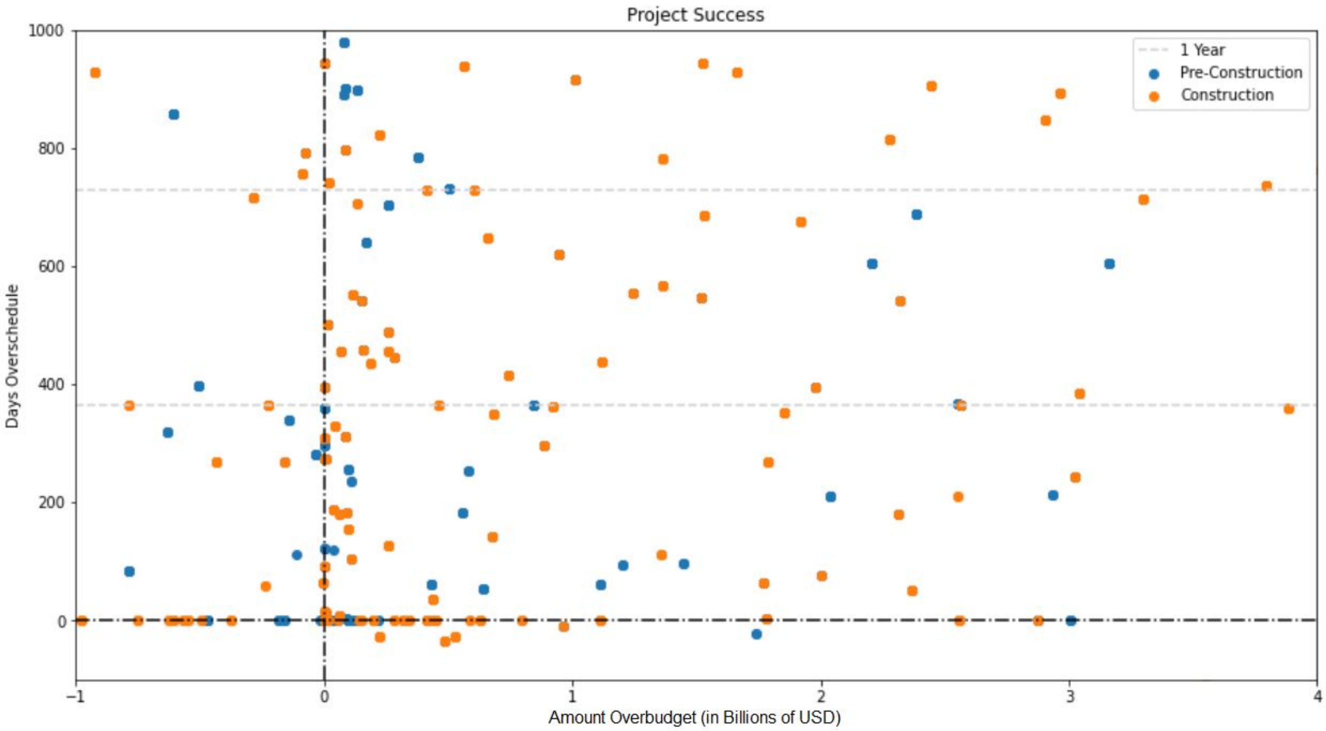
- We plotted histograms of the distributions for the **latest budget changes**, **total budget changes**, **latest schedule changes**, and **total schedule changes** variables. We noticed that all four histograms were centered around zero, which means that most project updates reflected no change from the original budget or schedule. We also noticed that all four histograms were skewed to the right. This indicates that there are more projects which are over budget/schedule than projects which are under budget/schedule. This observation is **in line with our experience and knowledge about capital project successes**.



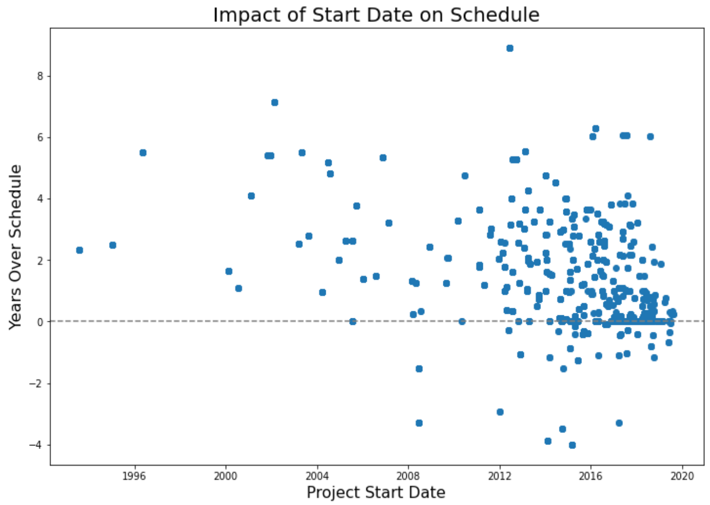


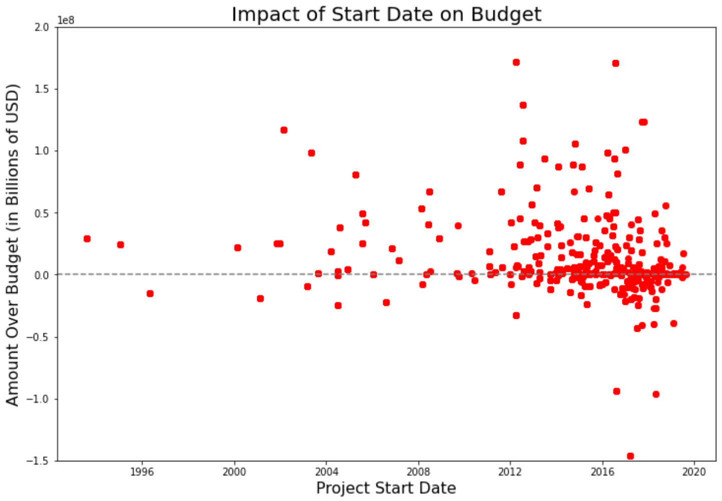


- We observe that the **largest number of capital projects** in the dataset are:
    - **managed by the DDC** (Department of Design and Construction), followed by the DOT (Department of Transportation), the DEP (Department of Environmental Protection), the EDC (Economic Development Corporation), and SCA (School Construction Authority). One of the areas our team would like to investigate are whether city agencies which are geared towards construction and development (ex. DDC, SCA) have a higher rate of project success than other agencies where design and construction are only a small part of the agency's mission (ex. FDNY, DSNY, etc.)
    - **in the Brooklyn** borough of New York City, followed by Queens and Manhattan. There are some capital projects which involve two or more boroughs, as well as other capital projects which lie outside of New York City.
    - **involve street and roadway infrastructure**, followed by water supply, wastewater treatment, sewers, and bridges.

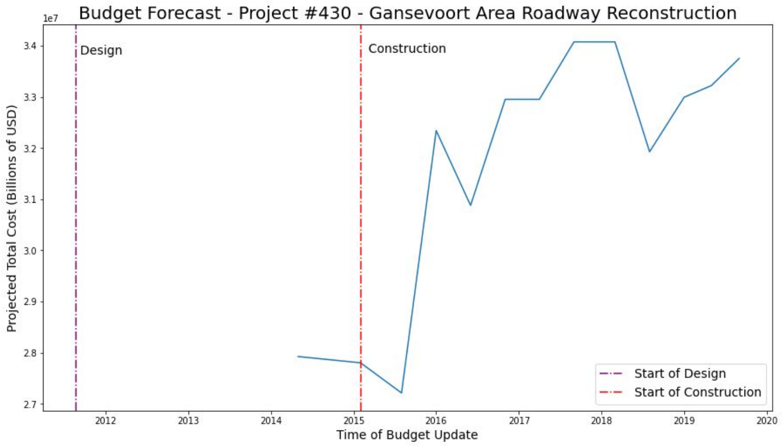


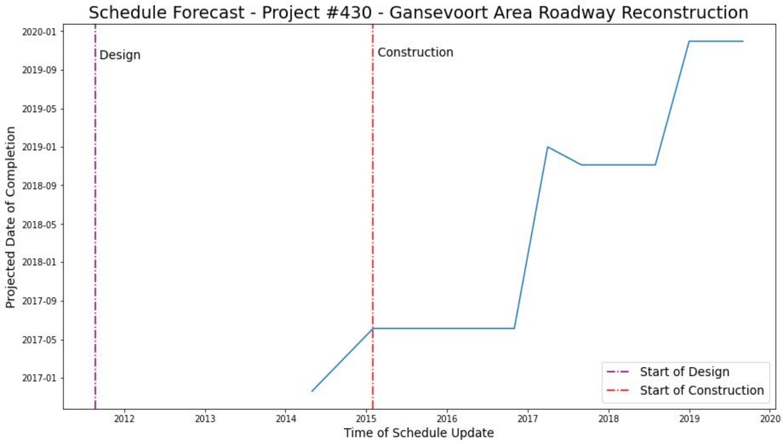
- We observe that the **majority of projects with serious temporal and cost issues are all in the construction phase.** We again note that due to vastly different scopes and types of project size, over budget and lateness are vastly different scales. This further supports the necessity of using a relative percentage error metric.





- We observe that time series will factor heavily into this project, as project start date, and elapsed time have a clear factor on budget and schedule. Perhaps obviously, plots above show the **longer a project runs the more difficult forecasting cost and budget becomes**.





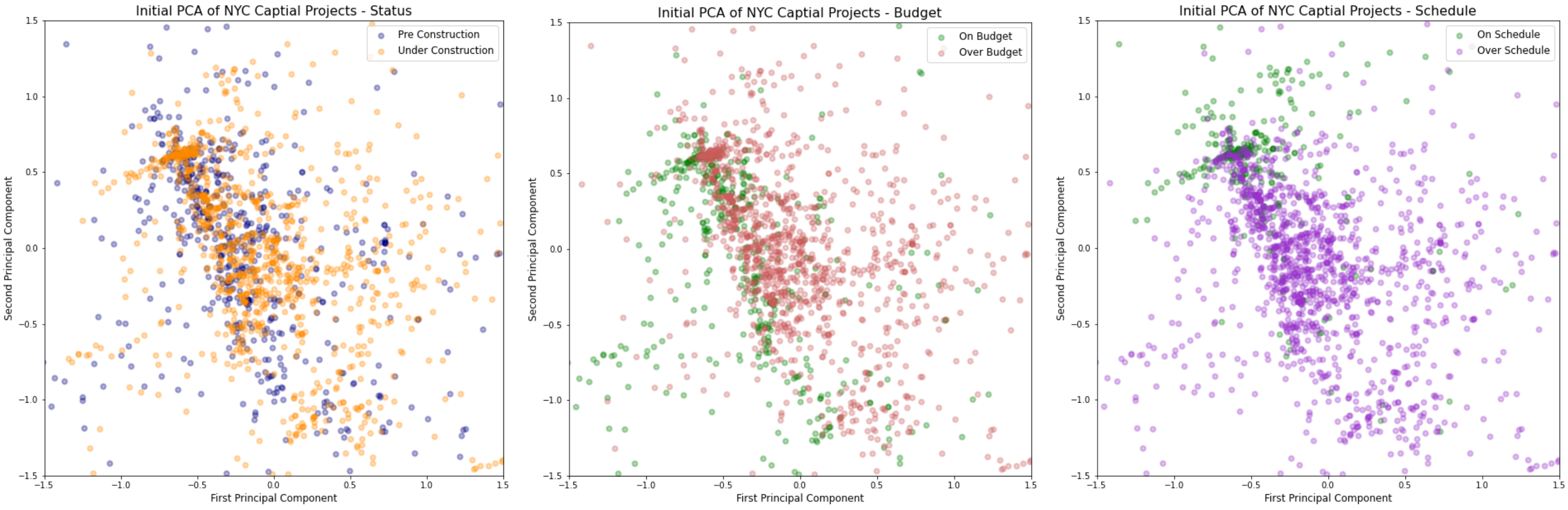
- To better understand the **time series data**, we spot analysed and compared 25 projects; Above is an *example* pair-plot for a roadway reconstruction project.

- A common theme is shown, that after the start of construction, cost and schedule rapidly increase. This demonstrates an issue we expect may be an issue in forecasting project success - A **lack of updated forecasts during the design phase of the project**, leading to an unrealistic expectation of time and cost.

- It is also of note, that many significant schedule updates occur about the time the last update indicated the project would be complete - this **poor intermittent updating also makes it difficult to anticipate changes forecast time and cost**.

- This time series data above implies a **slope/rate of cost and schedule change** at various phases that is of interest. The elapsed time at which this slope changes could be useful.

Principal Components Analysis - Clustering Feasibility



- To initially examine if there are sufficient descriptor variables to allow the data to be successfully separated by "success" metrics into clusters, **we performed dimensionality reduction on the full dataset**.

- We first investigated 'project phase', 'budget status' and 'schedule status' as cluster labels.

- Principal components analysis was only moderately successful, capturing just over half of the variance in the dataset (explained variance in the first two principal components was **31.67%** and **20.94%**).

- This shows the **dataset needs further data augmentation to properly cluster** (by success); Even by category, there's insufficient variance to separate, and draw clear class boundaries.

Summary

Exploratory Data Analysis shows that there is a wealth of time-series data available, but **further augmentation, tagging and processing is required** to extract its full value and produce a rich predictive model. Enhancement of time-series and text data and is a priority to properly cluster it.

Next Steps

As an <u>initial goal</u>, we want to **build a model to forecast a current project's absolute percentage error -** it is anticipated this will be **a GAM based model** (validated using ARIMA). We also look to categorize the project data by these metrics, by auto-encoding the project into a lower dimensional latent space (using auto-encoder neural networks), and labeling that latent space.

As stretch goals we look to predict a project success in as much detail as possible, if it is **possible to forecast the running future of a project**, and evaluate different trajectories each project may take (given certain choices). To that end we will evaluate and attempt to apply Recurrent neural networks (RNN), variational-auto-encoding (VAE) and reinforcement learning.

Using **VAE** would allow us to sample a hypothetical new project of a specific type and schedule/budget status from existing projects- it could provide a detailed forecast, that is especially **useful for projects with size and scope that does not match any existing project**. A challenge to this is that there may likely be insufficient variance in the data to sample from a continuous latent space generated by a VAE in a way that the resulting "theoretical project" is meaningful (the results may be highly biased).

Reinforcement learning would be tricky to apply, but provide enormous predictive power at any project state by recommending optimal value decision making. If we can successfully infer states and actions, we can re-evaluate existing projects by off-policy sampling - determining the value of alternate choices to find better ones. Additionally, if we can and derive transition probabilities by sampling from existing projects, then we can evaluate brand new projects - **simulating trajectories and selecting optimal policy for new projects** at each step.

Recurrent neural networks are likely to have the highest chance of success, specifically Gated Recurrent Units (GRU) at predicting time series data. We will initially attempt many-to-one models, that ingests past project series and **generates the next budget and schedule update**. Initial research indicates a time series bias needs to be applied and calibrated, as the forecasted values may become out of step from the actual ones (being consistently late). We may find neural networks prove to be too powerful to utilize on such a low dimensionality dataset.

References

Baucells, M., Grushka-Cockayne, Y., Hwang, W. (2019). The Effects of Mental Accounting on Project Performance.

Grushka-Cockayne, Yael. (2015). New York City Department of Parks and Recreation. Darden School Foundation UVA-QA-0815TN. Charlottesville, VA: Darden Business Publishing.

Lukas, J. A. (2012). How to make earned value work on your project. Paper presented at PMI® Global Congress 2012—North America, Vancouver, British Columbia, Canada. Newtown Square, PA: Project Management Institute.

Mayor's Office of Operations. (2019, November 15). Capital Projects: NYC Open Data. Retrieved April 13, 2020, from https://data.cityofnewyork.us/City-Government/Capital-Projects/n7gv-k5yt