Capstone Project
Pandas Express - Connor Haas, Richard Kim, Marek Kwasnica, Mohamad Sayed
Team self-assessment

1. After finishing your project, are you able to fulfill all objectives of your proposal?
Please list the steps from your proposal and respective achievement rates and shortcomings. If you achieved more than what you had proposed originally, please add it and describe your major achievements.

Steps/items	% comp.	Comments
Step 1: Setup Git, Tableau, SQL, compile Dataset	80	Dataset, Git and Tableau successfully implemented. Future work will be to integrate a Docker container with a MySQL database for ease data storage and retrieval.
Step 2: Data cleaning and Summary EDA	95	Dataset nicely saved to a csv file, cleaned and downsampled. Good characterization of missingness and data value types and value counts in data features. Could clean with some helper functions for easier reproducibility
Step 3: Finalized clean dataset and in depth EDA	90	Used visualizations to see features that had high correlation with admittance to the NICU. Cleaned and produced a full dataframe of all years by dropping certain features with high missingness.
Step 4: Tableau	75	Working dashboard, with most of desired visualizations and functionality
Step 5: Machine learning	95	Models trained and hyperparameters tuned. Models save and can predict given new info.
Step 6: Github & code cleanup	20	Consolidate and finalize notebooks and helper functions, finish readme file.
Additional item: Github project Management, Cookiecutter, Docker File	10	Kanban style project management on github was not as helpful or necessary as we anticipated and thus underutilized. Cookie cutter framework and docker files not fully integrated at this time.

2. What are the major obstacles your team has faced to fulfill your project objectives?
Please analyze it through several tangents.

Data collection and preparation	The large dataset took longer to extract and convert to csv than anticipated. Unclear sources of missingness that varied in level year to year in the data. (Eg. Initially wanted to model years 2010-2015, needed shift/drop years to work with more consistent datasets). Large number of features, inconsistent null values. Some features added/removed between years, unable to due time series analysis in our time frame for this.
Business insights generation	A secondary dataset of socioeconomic and geographic data would have been necessary to make the full use of our dataset. Needed to rely on outside academic and reporting services to justify economic motivation.
Proposal design	Making sure that we could justify our claims with the features available. Making sure we had correct data given difficulties in acquiring data. A large number of features made for a complicated interaction landscape.
Team organization and management	Lack of central meeting place led to some small technical issues and project direction decision-making taking longer to resolve than would have in person. Otherwise, no problems.
Technical strength/weakness	Unable to wrangle Docker and MySQL server to serve as foundation for notebooks and Tableau.Ended up just using csv files. Implementing Dashboard in Tableau entailed re-doing much of our Matplotlib visualizations Plotting skills were initially stronger in R than in Python, took a while to get a solid grasp of Matplotlib
Unforeseen issues	-Integration of Python code and Tableau; there is a TabPy library, but visualizations in Matplotlib do not port directly to Tableau -Was not clear how Tableau could have been a better used as collaboration tool

3. Please list the contributions of individual team members in the projects and their roles.
Write in % the overall contribution of each team member.

Team members	% contr.	Comments
Connor Haas	40	Data set parsing and prep, visualizations, random forest model machine learning. MVP
Richard Kim	20	Machine learning - random forest, and linear modeling; K-means clustering support pair-programming and debugging
Marek Kwasnica	20	Team Git monkey, project management. Data EDA and visualization, (Docker and SQL implementation), Storyboarding.
Mohamad Sayed	20	Data prep and cleaning, K means clustering, Tableau visualizations

4. In finishing your project, summarize what your team, both at an individual level and at a group level, have learned in the process.

- Learned to first look at how desired systems and tools will integrate before choosing development path (MK)
- Writing optimized helper functions without for-loops, and collaborating with multiple team members focusing on different sub-products of whole project (Team)
- Tele-collaboration and exploring different project organization techniques (git, kanban,shared google doc meeting notes, pair-programming, meeting scheduling frequency, task division and assignment) (Team)
- Matplotlib and Pandas (MK and CH)
- Random Forest (RK)

5. If the capstone project starts over again, what would you do differently to address the issues you have identified and encountered.

1. Devote a day or two to look over each proposed tool/system and more carefully consider feasibility and integration with data acquisition and project workflow demands to avoid bottlenecks.
2. Plan fewer daily full team check-ins in favor of more pair-programming
3. Not split up at the beginning to try to download and parse one year of data files. Having one dedicated sub-team to work through each file would have been faster , easier, and more consistent.

6. Can you think of other industries and topics on which you could apply the same technique and methodology you used in your capstone project?

1. A lot of the information we learned could be used in situations that have large datasets that need feature selection and sampling.
2. We did a lot of analysis on binary features and could target specific customers needs based on previous purchasing history.
3. Using classification modeling and time series analysis can help predict trends in finance and other risk based fields.

7. Please list the names of instructors and TAs you have discussed with for your project and estimate the amount of time you have spent with them.

Staff members	Time	Comments
Instructor 1: Thomas	4+ hours	Couldn't and wouldn't have done it without you ;). Very helpful answering questions and keeping us on track for getting an end result and telling our unique story.
TA 1: Sam Audino	1/2 hour	Questions regarding SQL, Tableau, Docker