HemaSource

HemaSource Analytics Project

Brooks Walls, Jacob Sanders, and Caleb Pan

Kayla Andersen with HemaSource

May - June 2020

Table of Contents

Introduction

HemaSource is the primary medical supplies provider for most of the plasma donation clinics throughout the US. Their business revenue is strongly correlated with the number of donation centers and the donation volume (number of procedures) from each of those centers. HemaSource currently lacks substantial information around each of the centers and their relative customer (donor) base markets, making it difficult to compare and contrast centers and their performance.

The primary goals for our field session project included providing HemaSource with valuable analytics centered around individual donation centers, as well as a forecast for future donations. Our analytics focused on three main areas: donation trends, demographics, and the effects of COVID-19 on donation centers and their donation volume. To accomplish these goals, we used HemaSource's donation history data and center information as well as census and Zillow data to get the best picture of each center's performance and donor pool.

The final product will consist of presentable analytic products for the three areas of focus, as well as reusable code that can predict future donations for centers. The data warehouse built for these tasks should also be able to integrate with HemaSource's existing data lake and be well documented to allow future use.

Requirements

Functional Requirements

Database Product

1. Collect external third-party data to be used with internal HemaSource data
2. All third-party data available in S3-based cloud
3. Code-based ETL pipelines
4. Build a relational database (bridge internal and external data)
5. Data dictionary for the database
6. Auxiliary:
    a. Feature engineering tables
    b. Unit testing for reusable ETL pipelines

Analytics Product

7. Predictive model input
8. Reusable predictive analytic models (based in Python/scikit-learn)
9. Business Intelligence Tables (short and long-term, digestible reports)
    a. Examples: top five locations for centers, projected gains/losses
10. Data Visualization Products of the above
11. Internal presentation to HemaSource

Non-Functional Requirements

1. Use Conda environment
2. Use AWS (S3, RDS, EC2) for cloud computing and storage
3. Programming Language: Python 3.7
    a. pandas
    b. sqlalchemy
    c. scikit-learn
4. GitHub
    a. Project repository: data engineering/data science workshop space
    b. Utility repository: reusable code/implementations
5. Assessable presentation medium

System Architecture

Data Warehouse

One aspect of our final project design is the data warehouse that we built using individual tables that we received from Hemasource and third-party sources. The two tables that Hemasource provided contained information about donation centers and donation history. The three tables that we acquired from third-party sources included CBSA, property value, and census information.

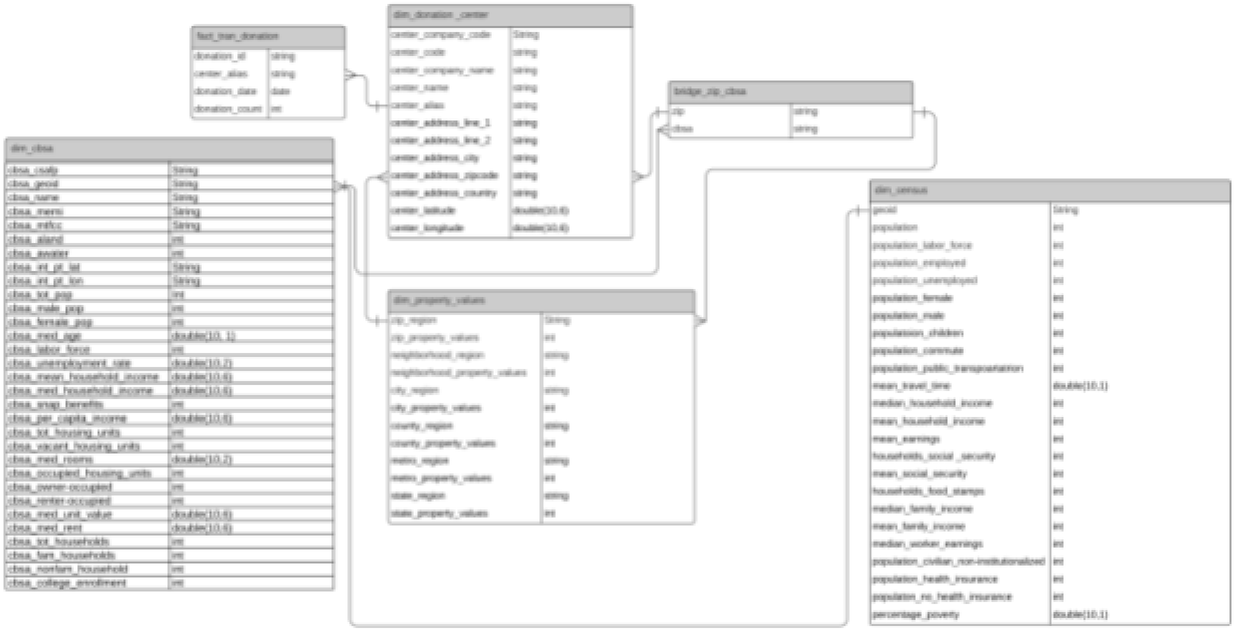
fact_tran_donation
donation_id string
center_alias string
donation_date date
donation_count int
dim_donation _center
center_company_code String
center_code string
center_company_name string
center_name string
center_alias string
center_address_line_1 string
center_address_line_2 string
center_address_city string
center_address_zipcode string
center_address_country string
center_latitude double(10,6)
center_longitude double(10,6)
bridge_zip_cbsa
zip string
cbsa string
dim_cbsa
cbsa_csafp String
cbsa_geoid String
cbsa_name String
cbsa_memi String
cbsa_mtfcc String
cbsa_aland int
cbsa_awater int
cbsa_int_pt_lat String
cbsa_int_pt_lon String
cbsa_tot_pop int
cbsa_male_pop int
cbsa_female_pop int
cbsa_med_age double(10, 1)
cbsa_labor_force int
cbsa_unemployment_rate double(10,2)
cbsa_mean_household_income double(10,6)
cbsa_med_household_income double(10,6)
cbsa_snap_benefits int
cbsa_per_capita_income double(10,6)
cbsa_tot_housing_units int
cbsa_vacant_housing_units int
cbsa_med_rooms double(10,2)
cbsa_occupied_housing_units int
cbsa_owner-occupied int
cbsa_renter-occupied int
cbsa_med_unit_value double(10,6)
cbsa_med_rent double(10,6)
cbsa_tot_households int
cbsa_fam_households int
cbsa_nonfam_household int
cbsa_college_enrollment int
dim_property_values
zip_region String
zip_property_values int
neighborhood_region string
neighborhood_property_values int
city_region string
city_property_values int
county_region string
county_property_values int
metro_region string
metro_property_values int
state_region string
state_property_values int
dim_census
geoid String
population int
population_labor_force int
population_employed int
population_unemployed int
population_female int
population_male int
populaiosn_children int
population_commute int
population_public_transpoartation int
mean_travel_time double(10,1)
median_household_income int
mean_household_income int
mean_earnings int
households_social _security int
mean_social_security int
households_food_stamps int
median_family_income int
mean_family_income int
median_worker_earnings int
population_civilian_non-institutionalized int
population_health_insurance int
population_no_health_insurance int
percentage_poverty double(10,1)

Figure 1

Above is our entity relationship diagram (Figure 1) for the data warehouse we built (full-sized version in the appendix). This data warehouse utilizes star schema where one fact table references the surrounding dimension tables. Our fact table contains daily donation history for every plasma donation center that HemaSource supplies going back to 2011.

Using the information in the `fact_tran_donation` table we can connect to various dimension tables that contain different information that may be valuable for different types of analytics. `dim_donation_center` contains all of the relevant information for each donation center, including the center address, name, hkey (hash key), parent company, and latitude and longitude. Using this information and our bridge table that converts zip codes to Core Based Statistical Area (CBSA) codes we can pull census data about the donation center's demographics. CBSAs are used by the census to donate areas made up of multiple zip codes that act as one city. By using our dim_CBSA table and the census data we can pull a copious amount of data on a donation center's surrounding area and donor pool.

The benefit of using star schema is that we can use simple joins to pull many features at once, and still have relatively fast fetch times. Since we are not processing transactions we do not have to utilize the fastest database schemas, and thus we do not have to worry about data redundancy.

Workflow & Process

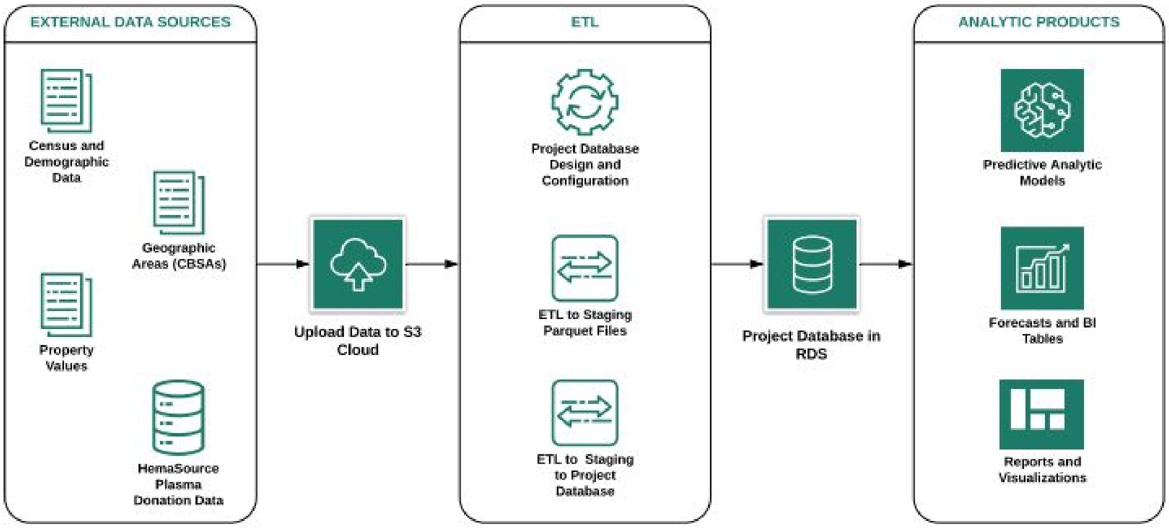
EXTERNAL DATA SOURCES
Census and Demographic Data
Geographic Areas (CBSAs)
Property Values
HemaSource Plasma Donation Data
Upload Data to S3 Cloud
ETL
Project Database Design and Configuration
ETL to Staging Parquet Files
ETL to Staging to Project Database
Project Database in RDS
ANALYTIC PRODUCTS
Predictive Analytic Models
Forecasts and BI Tables
Reports and Visualizations

Figure 2

Figure 2 shows our diagram for the entire project and its workflow. The initial section covers the sources of our data that we would pull into our new data warehouse. From there we uploaded the source CSVs and parquets to AWS' S3 service where our whole team could access the files.

The next section covers our ETL, the process where we Extract, Transform, and Load our data. In this section of the project we designed our database schema, planned any necessary data engineering to clean up the source data, uploaded clean parquets to S3 once again, and then finally took the files in our staging bucket in S3 and uploaded it to our AWS RDS which is a relational database on the cloud platform.

The final section was the analytics and forecasting that utilized the data we pulled and uploaded. These analytics focused on donor demographics, donation trends, and the effects of COVID-19. The analytics were presented in multiple ways, including Jupyter notebooks, interactive HTML files, and presentation slides. We also used our database to build a random forest model to predict future donations for individual centers. Finally, we supplied HemaSource with a cosine similarity model that was used to identify the best CBSA's to build future donation centers in.

All of the models we utilized were fed in data through SQL queries that accessed our Postgres database. These queries were read into the Python using SQLalchemy and pandas dataframes. Once the data was in dataframes we would normalize them if necessary and feed them into the model where the analysis could be done.

Technical Design

COVID-19 Analytics

One interesting area of our project was the analytics we did that looked at COVID-19's effects on donation centers and their daily donation volume. This is especially relevant and gave our team a new perspective on the situation. The figure below shows total donations in millions from all centers for every month of each year for which data is available.

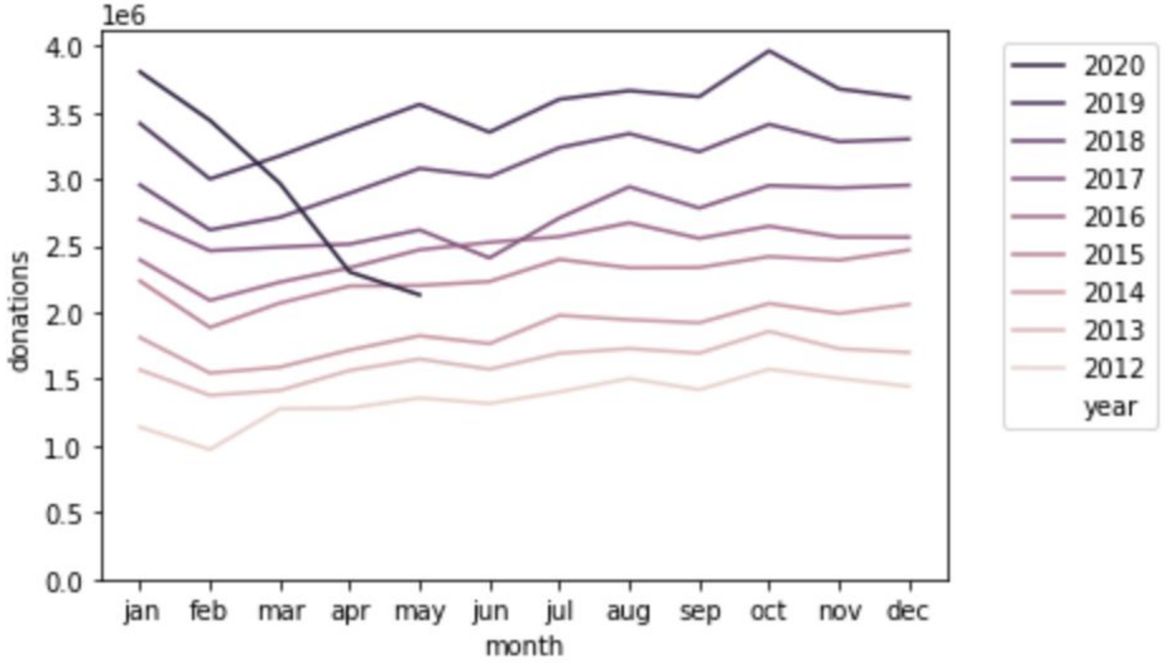

Figure 3

From figure 3, we can see that each year steadily increases likely due to new centers opening and populations increasing. However, the trend for 2020 is unprecedented as donations have decreased every month. This is no doubt due to the global pandemic.

Going into the analytics we assumed we would be able to identify which centers, CBSA's, and States are hardest hit from COVID-19 and which were not affected very much if at all. However, once we began looking at the data it was clear that no area or center had not been affected by COVID-19. Below is a screenshot from an interactive map we made utilizing ipyleaflet in Python. Each dot represents a donation center, and the color signifies the percent drop in Donations from April 2019 to April 2020.

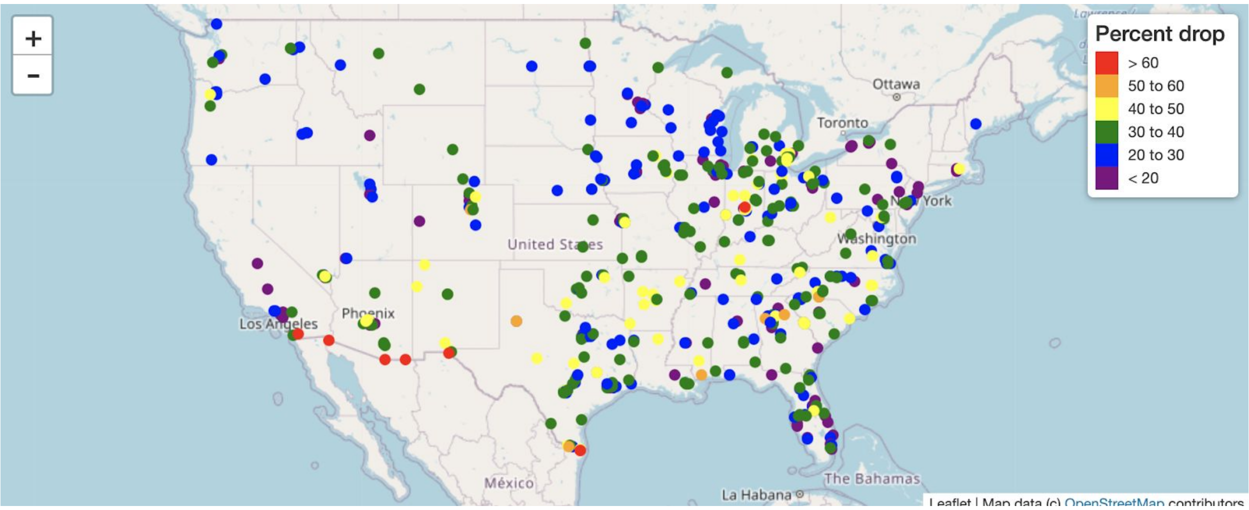

Figure 4

In figure 4, we can see that the hardest hit area is the Mexican border. Only one red dot can be found elsewhere. Before COVID-19, these centers had fairly high daily donations consistently thanks in part to donors who would cross the border to donate. Below is a graph of one of these "red dot" center's daily donations in 2019 and 2020.

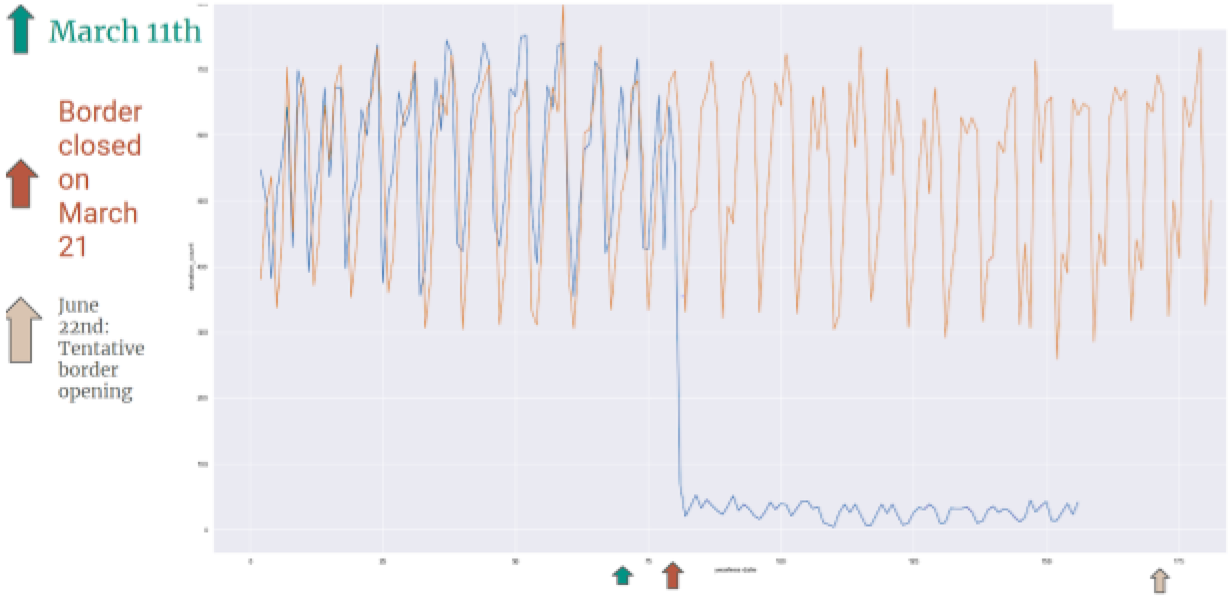

Figure 5

In figure 5 above the center’s 2019 daily donations are graphed in orange and 2020’s daily donations are graphed in blue. Before COVID-19, daily donations are extremely consistent year-round and year to year. The arrows at the bottom of the figure denote key events in 2020. On March 11th, 2020 COVID-19 was declared a pandemic, and on March 21st, 2020 the US/Mexico border was closed. The gap in donations did not drop, only once the border was closed did daily donations plummet. This behavior of donations is how all near-border donation centers behaved this year. Based on the lack of drop-in daily donations between March 11th to March 21st we assume that without social distancing regulations and a complete opening of the border these centers will quickly return to their usual donation volume. However, if border crossings are limited or the centers have social distancing regulations then they most likely will see donations increase but not return to their usual daily amount.

Original Theories & Model Results

One of the major objectives was to find the demographic information of the surrounding areas for each center. Since HemaSource had little knowledge and only speculation, a data-supported theory would greatly help in decisions of where to open new centers.

Since plasma donation is a source of additional disposal income (for any demographic), we and our client came up with a theory that areas with less economic wealth were more likely to donate. Therefore, we expected a decrease in average daily donations as an area became more wealthy. However, our models differed from the expectations of our theory.

There are many different measurements of the economic wealth of a geographical area. Among the measures that we implemented include mean family income, median worker earnings, and percentage of families on food stamps. Below is a plot that illustrates the relationship between mean family income and average daily donations.

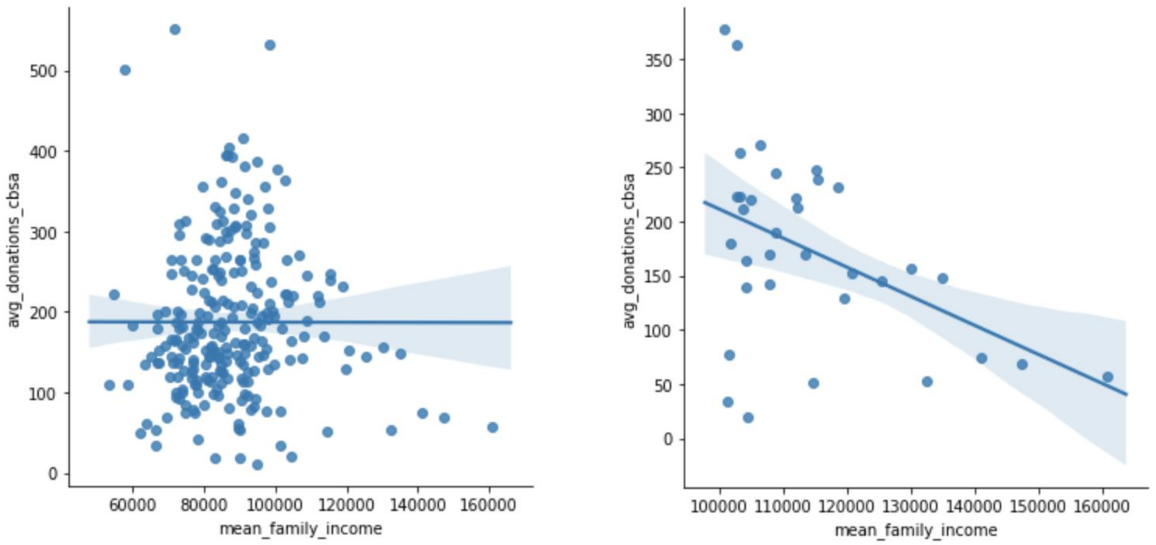

Figure 6

Figure 6 (left) shows a plot that illustrates the relationship between the mean family income of a CBSA and the average daily donations from the centers. As we can see, there is no obvious relationship between mean family income and average donations. However, Figure 6 (right) is the same plot zoomed in to CBSA's with a mean family income of $100,000 and up. We can see a clear downward trend with the "richest" areas. This confirms our hypothesis that areas with less economic wealth are more likely to donate, but we can only apply this to roughly the wealthiest 10% of areas.

Another initial theory we had involved cannibalization. We predicted that when a donation center opens up, it will steal, or cannibalize from the donor base of nearby centers. The figure below shows the relationship between the distance between a newly opened center and one nearby, and the change in donations for the nearby center 30 days before and after the new one opens. The left shows the percentage change while the right shows the change in the number of donations.

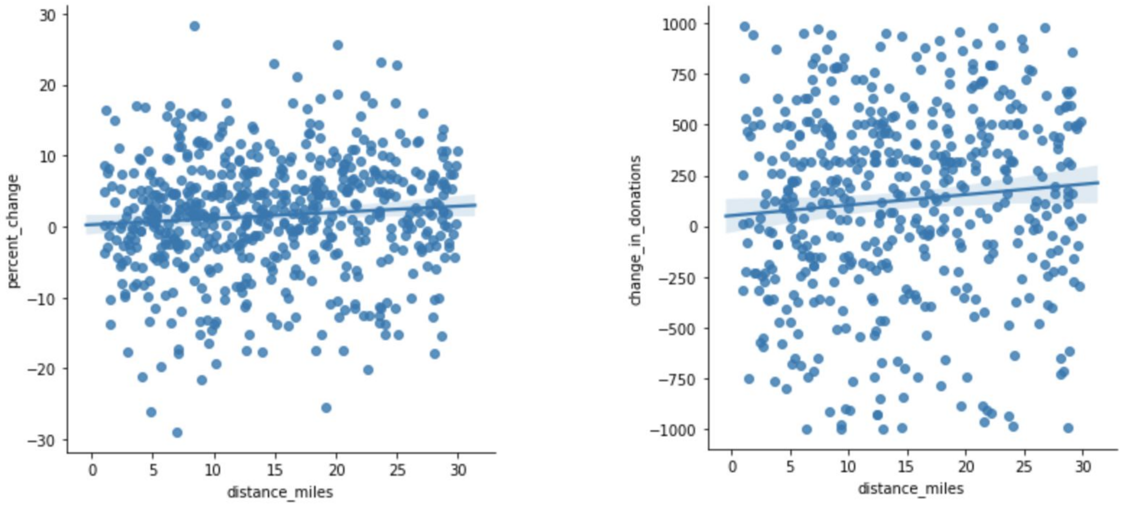

Figure 7

From figure 7, we can see that a center opening nearby has minimal effect when disregarding other variables. However, a slight upward trend can be noted as the distance increases. Based on the left plot, for a center 30 miles away from a new center, we can expect to see a 3 to 4 percent increase in donations. For a center within 2 miles of a new center, we can expect to see a -1 to 1 percent increase in donations. However, more analysis is necessary to be able to determine definitively.

Quality Assurance

Code Quality

Backlog

Using the backlog we can continuously ensure we are meeting our client's requests for the project. While using a general backlog to track the overall progress of the project we also use a weekly sprint backlog to track in-depth stories. This weekly backlog allows us to communicate with the client and have records of the requirements for each aspect of the project.

Repository Checks

Our team and our client share two Git repositories: utilities and projects. We have two repositories for organization. As we continue to build reusable code (utilities) and create our client deliverables (projects), our frequency of pushing onto the repositories has steadily increased. Because of the high activity, we practice review and merge requests for all code commits. This allows our client and team members to evaluate code contributions concerning current progress before accepting.

Accuracy Scores

Once our predictive models are built we can designate previous months in the donation history table as test data. Using this test data we can get accuracy scores for our models to ensure they can be used to give a general idea of future donations. Ideally, our models will have around or above 80% accuracy to ensure good predictions without overfitting the training data.

Team and Client Communication

Daily Stand-Up Meetings

Every weekday at 11:00 am, our team meets to bring everyone up to date on the information that is vital for coordination: each team member briefly describes any completed contributions and any obstacles or difficulties experienced. There is then a discussion of items needed to be discussed with our client and/or advisor. The meeting concludes with each team member describing what they are working on for the day. This meeting is a typical duration of 15-20 minutes.

Group Communication

Outside of Daily Stand-Up Meetings, we communicate very frequently through group chat if any issues arise, help is requested, or suggestions thought of to be shared. Additionally, Zoom meetings among team members are scheduled for more in-depth collaboration. Sharing code and examples are commonplace to bridge gaps in knowledge and skill.

Client Meetings

Our team has scheduled three meetings a week with our client: Mondays we have sprint planning, Wednesday we have office hours, and Friday we have demonstrations for our client. This follows the cycle of initializing, developing, and review. We typically schedule additional meetings as well (both by our client and our team) to receive additional mentorship on technologies and skills we are unsure of.

Results

Hypothesis & Models

Economic Demographics: our original hypothesis was donation centers in areas with lower economic wealth (income, property values, employment, receiving benefits) are likely to have higher donation averages. Our reasoning was individuals with less wealth would be more likely to donate to gain additional income. However, different factors had different results with our models compared to our expectations.

- **Benefits**: there seem to be no strong correlations for households on food stamps or the amount of social security received, but there seems to be a negative correlation with donations and areas with more households receiving social security
- **Income**: our models of income (worker, household, family) provided no clear trend except for the wealthiest roughly 10% of areas. In the wealthiest areas, there is a negative correlation between donations and income.
- **Employment/Labor Force**: there seems to be a strong positive correlation between employment and donations (the more a population is employed, the more donations are received). The inverse is also seen with unemployment.
- **Children**: there seems to be a positive correlation with donations and the number of households with children (under 6, 6-17, and combined) in an area.
- **Transportation**: there is a strong positive correlation with donations and a population that commutes to work. We observed the same for mean travel time (to work).
- **Property Values**: there seems to be no correlation with property values, although there is notable clustering of centers being located in areas with low property values.

From this, a new hypothesis can be formed based on our observations with data. Donation centers will have higher donations in areas with:

- Higher employment and income: while this may seem counterintuitive, while unemployed, the main objective is finding a job for sustainable income. Donating blood plasma is not a job and an esoteric source of income in comparison (probably not crossing a job-seeking individual's mind). Furthermore, donating involves time, energy, and transportation that an unemployed individual may not consider worth the investment compared to job-seeking.

- More households with children: this might be explained as raising children is expensive, which may be a higher incentive to donate for additional income.
- More workers who commute to work and who commute longer: this might be explained as workers who are comfortable with regular/longer travel are more likely to go out of their way to a donation center (as most donations occur on weekends when people are not working). Furthermore, this is supported by the implication workers have their reliable means for transportation.

COVID-19 Analytics Results

One area our team focused our analytics on was the effects of COVID-19 on centers and their daily donations. To do this, we utilized the supplied donation history table from HemaSource, as well as CBSA data and donation center data. The main areas we focused on was identifying the hardest-hit centers and CBSAs, and identifying centers or CBSA's that may be rebounding from the effects of COVID-19.

Overall we found that all Core Based Statistical areas have experienced negative effects on their daily donations due to COVID-19. The biggest variance came in identifying when different areas experienced their drops in donations and what caused it. One of the most interesting areas to look at were the centers that are located on the border. These centers did not experience any drop when COVID-19 was declared a pandemic on March 11; however, on March 21 when the border was closed these centers experienced 90% drops in their daily donations which have not gone up since. It can be expected once the border does open up again then these centers will return to their 2019 averages unless they experience social distancing regulations.

While there are about 5 centers that have April and May Daily donations higher than their March donations, there are no centers that were open all of 2019 that have current daily donations equal to or greater than their 2019 equivalent daily donations.

All of this shows us that the effects of COVID-19 have affected centers heavily and these centers will most likely take all of 2020 to recover.

Unimplemented Features

- Use of GIS/Tiger files
    - These files would allow us to easily display the location of CBSAs and their donation centers; however, we were able to plot centers on a map without these files.
- ETL for shapefiles
    - Since shapefiles require around 7 individual files for each shape we would require complicated ETL to ensure we have easy access to all the files and the data is cleaned properly.
- Analysis of donor base cannibalization
    - If we had more time we could further identify how much of an existing donor pool is taken by new centers opening near an existing center.

Future Work

- Stimulus Checks effect on donations
    - Since most people's motivation to donate plasma is the payment they receive for donating, HemaSource believes the stimulus checks could drop daily donations by a decent amount. This is because the $1200 amount is equivalent to donating multiple times a week for an entire month. These analytics could also cover the effects of other policies such as increased unemployment pay, increased minimum wage, and rent control.
- Type of donation center for an area
    - Since HemaSource supplies multiple Plasma donation companies that own multiple centers they have a wide variety in types of centers and their associated donation history. These different centers can vary from stand-alone buildings that emulate a hospital, to a business in a strip mall. With this data, HemaSource could begin to identify the best areas for each type of center.
- Social Distancing Regulations
    - If HemaSource can identify different areas' requirements for social distancing, especially if the quantity a center can process at a time is restricted, then this data can be used to fine-tune prediction models as restrictions are in place and slowly taken away.
- College's decreasing/increasing tuition effect on donations

    - As colleges around the country face unique economic challenges many schools are changing the price of tuition. This could have effects on college town donation centers and their daily donations. These changes would be tracked and used to identify how much the cost of schools affects donations at specific centers.
- Hyper-Seasonality
    - Depending on a center’s location they may have constant daily donations throughout the year, or they may experience spikes and falls in daily donations throughout the year. If the centers with variance in their daily donations can be identified then they could use specific models to predict their future donations.

Lessons Learned

Our team was rather inexperienced with ETL pipelines. As we glanced through our data we realized it wasn't as simple as run and upload. Some of the data was incomplete (ex: zipcode missing a beginning digit) so we would need to repair or filter out hazardous data in our ETLs. We also expanded our data engineering tool belt using AWS and S3.

We became experienced with the data science applications of Python, particularly with Pandas and Seaborn. We became familiar with data manipulation and modeling (determining what to include, exclude, timeframe, etc).

The sheer size of the data we analyzed provided challenges. We became skilled at writing complex SQL queries to snatch specific, usable data. Because we gathered so much data, much of it was messy and some of it irrelevant. Definitive trends were difficult to spot, and we learned that further analysis and “massaging” of the data (if we had more time) is necessary to confirm many of the trends we noticed.

There is an axiom that data scientists spend 80% of their time “data wrangling”(identifying, collecting, cleaning, aggregating, etc.) as opposed to data modeling and machine learning (see appendix for reference). We spent at least 80% of the time wrangling with our millions of data points as we are not as experienced as professional data scientists.

We learned how to interpret, theorize, and collaborate as data scientists (hypothesis, the reasoning behind trends, bouncing ideas off each other on what and how to analyze, etc.).

The importance of normalizing feature vectors and their data when using methods such as cosine similarity to identify similar donor populations.

Appendices

Figure 1:

https://drive.google.com/file/d/1pa3LE4krJF2L4dyvXhjOfItCqzs45QKI/view?usp=sharing

Figures 3, 4, 6 and 7 with additional explanation:

https://docs.google.com/presentation/d/1pOTADWIa3ZvG5FF4evZtQkMc87khoxEv__MqSX9Cw0Y/edit#slide=id.g889e4ec2b9_1_39

CBSA:

A CBSA is a Core Based Statistical Area that is used by the US census to denote areas centered around urban areas. These are made up of one or more counties surrounding urban centers with populations of at least 10,000 people. These help our analytics since near donation centers people often travel between counties and by only looking at zip codes we may not be able to capture the full demographics of a donor pool.

Odd Shaped Graphs:

You may notice that when looking at plots of daily donations for centers, CBSA's or States they have a roller coaster shape to them where they have frequent peaks and valleys. This is caused by the tendency of donors to visit the centers on weekends. Virtually all centers experience peak donations for a week on the weekends creating the peaks. Since there is less time to travel to a center during the week often donations are noticeably lower. This shape can be seen in figure 3 on page 6.

`generate_graphs` Usage Instructions:

This notebook can be used to quickly generate plots of a State, CBSA, or Center's 2019 and 2020 daily donations. These are especially valuable when looking at the effects of Covid-19 and how certain areas are recovering. The notebook starts at the state level and zooms into individual centers. To change what is looked up simply assign the given variable for each section the identifier for the State, CBSA, or center. Before each section, there is a lookup table that can be used to find information that will allow you to zoom in your analysis. This includes every CBSA in a state that has a center and all centers in a given CBSA. The best workflow for identifying a center would be to first gather the list of all CBSAs in a state and their number of centers, then from there use the provided CBSA GeoId, the unique identifier, and look up a table of all the centers in this CBSA. From there you can utilize the name of the donation center to generate the Center's 2019 vs 2020 daily donations plot.

Data Science Axiom:
https://blog.timextender.com/reversing-the-80-20-rule-in-data-wrangling#:~:text=Based%20on%20the%20results%20from,data%20modeling%20and%20machine%20learning