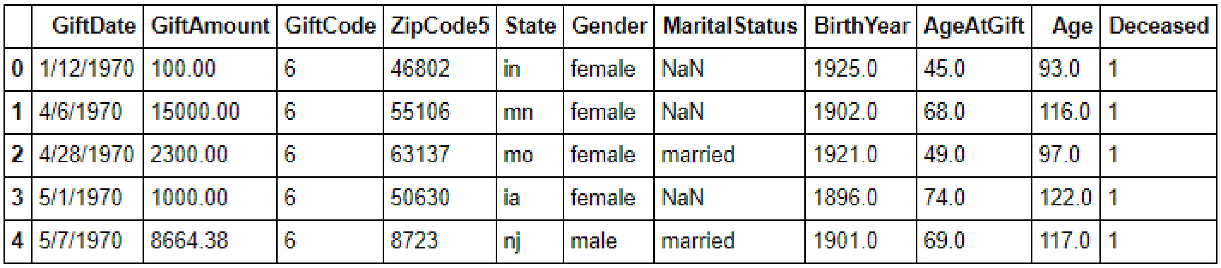
Harrison Lu
Michelle Xu
Meghan Woodruff

Midterm Progress Report

Each year, the Lutheran Church- Missouri Synod Foundation receives a certain number of irrevocable gifts. To predict the amount that the Church will be receiving on an annual basis is helpful in assisting the Church plan financial operations. We have tasked ourselves with finding an accurate and representative model of the annual amount in donations based on factors such as trends amongst people from different zip codes, ages, marital statuses, and other personal factors, as well yearly fiscal data, birth rate, and unemployment.

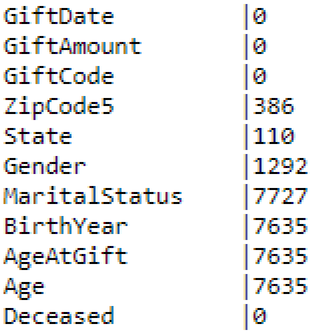
We begin by reviewing the format of our data. Data was collected when donations were made, and whatever personal information the individual felt comfortable divulging was recorded. The data included gift date, amount, gift code (which refers to the type of gift that was given), zip code of the location in which the gift was given, state, gender, marital status of the donor, birth day, age of donor when gift was received, current age, and finally, whether the individual is deceased. Figure 1 shows five rows of sample data.



	GiftDate	GiftAmount	GiftCode	ZipCode5	State	Gender	MaritalStatus	BirthYear	AgeAtGift	Age	Deceased
0	1/12/1970	100.00	6	46802	in	female	NaN	1925.0	45.0	93.0	1
1	4/6/1970	15000.00	6	55106	mn	female	NaN	1902.0	68.0	116.0	1
2	4/28/1970	2300.00	6	63137	mo	female	married	1921.0	49.0	97.0	1
3	5/1/1970	1000.00	6	50630	ia	female	NaN	1896.0	74.0	122.0	1
4	5/7/1970	8664.38	6	8723	nj	male	married	1901.0	69.0	117.0	1

Figure 1. Five rows of sample data.

We have a total number of 26,897 data points, some of which have missing attributes. While certain attributes such as gift date and gift amount do not have any missing rows, other attributes such as marital status, birth year, age at gift, and current age have over a quarter of the rows missing, as can be seen in Figure 2. When creating our model, the missing data is considered and processed accordingly, depending on the model.



```
GiftDate         |0
GiftAmount       |0
GiftCode         |0
ZipCode5         |386
State            |110
Gender           |1292
MaritalStatus    |7727
BirthYear        |7635
AgeAtGift        |7635
Age              |7635
Deceased         |0
```

Figure 2. Written next to each attribute is the number of missing data points of that attribute.

We ultimately decided to run our data mining algorithms in Python for two main reasons. Primarily, Python is commonly used for data science and already has a multitude of frameworks that simplify data science through better data organization and variable management. Secondly, our team members are already familiar with Python.

We began our data analysis by analyzing several independent variables. The most obvious and trivial would be to determine the frequency of each amount of donation over the entire dataset. Since amount donated is not a discrete value, we designated a series of evenly spaced buckets where the range of each bucket is 500. These buckets spanned from the lowest recorded donated amount, which was $0, to the highest amount donated, which was $5,402,615. This resulted in approximately 10,000 buckets. We assigned each donation to a bucket if it existed within the bucket's range, and then counted the number of elements in each bucket. We plotted our results on a log-log scale.

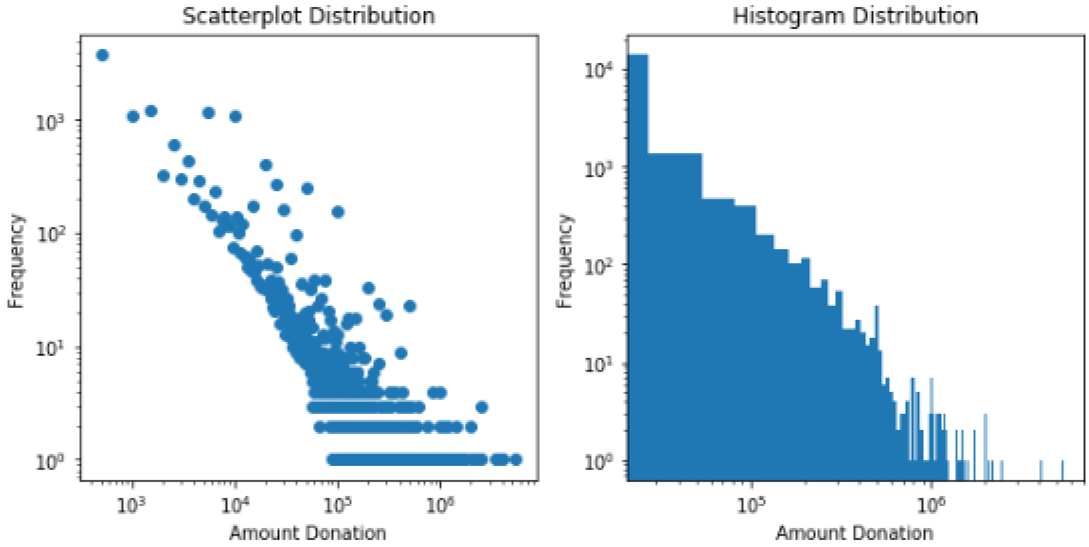


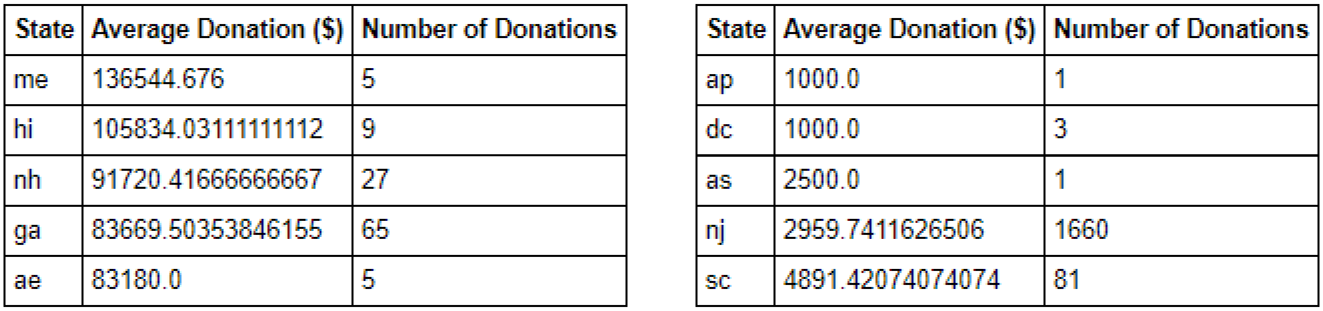
Figure 3. Frequency of donation amounts.

From the scatterplot in Figure 3, it is evident that, while the majority of people donate within $10,000, there are potentially around two hundred people who donate well above $100,000 and probably less than 20 who contribute amounts into the millions. This pattern is even more evident in the histogram, where the outliers are included in the bars. We can see that the frequency of donations versus the amount donated follows a power-law distribution, which is to be expected.

We were then curious about the discrepancy between the average amount that females donate and the average amount that males donate. (Note that I say females and males instead of women and men, because some donors are less than 10 years old.) Our results showed that females donated on average $29,625.99, while males donated on average $31,557.35. While these results suggest that males donate on average almost $2,000 more than females do, the standard deviation for each average was 128,406.85 and 123,429.65, respectively. This indicates that we have humongous variance in our data, and that averages are not necessarily a good representation of population. Nevertheless, the median donation for females was $4,025.04, and for males was $5000- indicating that there is a trend where males tend to donate more than females do, although a specific value is difficult to determine.

We wanted to analyze a few more independent variables before developing our model, so that we have an idea of how independent variables effect on donation amount. We found the five states with the highest average donation, as well as the five states with the lowest average donation, shown on the following page. It becomes apparent that the states with the highest average donation have, in general, very few donors, indicating there exists a small group of people who donate very large amounts. It is also possible that the far-right outlier from our scatterplot in Figure 3 resides in one of these states, skewing our results far to the right.

It was difficult to find a correlation within the lowest average states. The top three states had very few donors, but the fourth lowest state had well over 1500 donors.

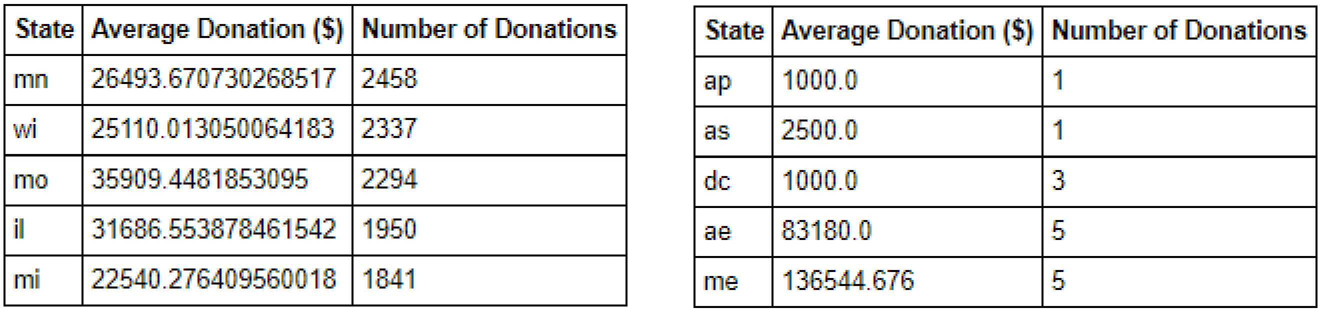


State	Average Donation ($)	Number of Donations
me	136544.676	5
hi	105834.03111111112	9
nh	91720.41666666667	27
ga	83669.50353846155	65
ae	83180.0	5

State	Average Donation ($)	Number of Donations
ap	1000.0	1
dc	1000.0	3
as	2500.0	1
nj	2959.7411626506	1660
sc	4891.42074074074	81

Figure 4. Left: Top 5 highest averaging states. Right: Top 5 lowest averaging states.

We also pulled the top 5 states with the most donations, as well as the top 5 states with the least donors. In Figure 5, the states with the greatest number of donations all have a relatively high annual gift value. However, the top 5 states with the lowest number of donations all have extremely low annual gift values. Combining the data derived from the top 5 lowest average donation in Figure 4 and top 5 lowest donations in Figure 5, we can begin to glean a correlation between a low number of donations and a low average donation.

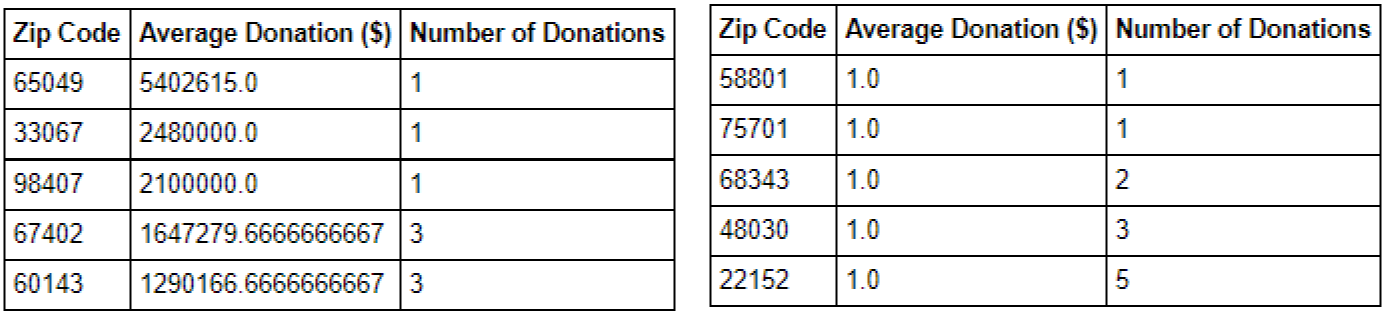


State	Average Donation ($)	Number of Donations
mn	26493.670730268517	2458
wi	25110.013050064183	2337
mo	35909.4481853095	2294
il	31686.553878461542	1950
mi	22540.276409560018	1841

State	Average Donation ($)	Number of Donations
ap	1000.0	1
as	2500.0	1
dc	1000.0	3
ae	83180.0	5
me	136544.676	5

Figure 5. Left: Top 5 highest states with greatest number of donations. Right: Top 5 states with least number of donations.

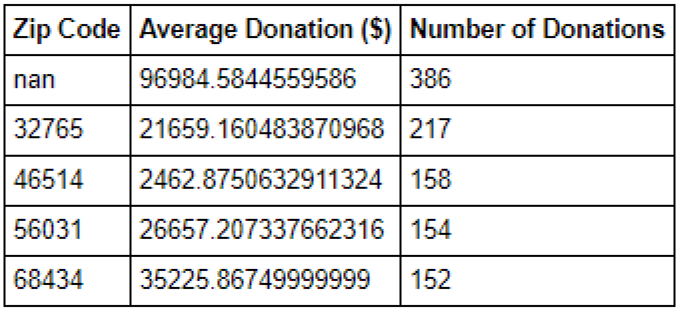
A similar pattern can be observed within zip codes. Note that because of a certain way in which the method that was used to calculate the average donation was implemented, an average donation of $1 actually indicates an average donation of $0, since for some reason, donations of amount $0 were included in the original dataset. Additionally, zip code "nan" indicates that no zip code, or an invalid zip code, was provided. The method used grouped all of those values together.



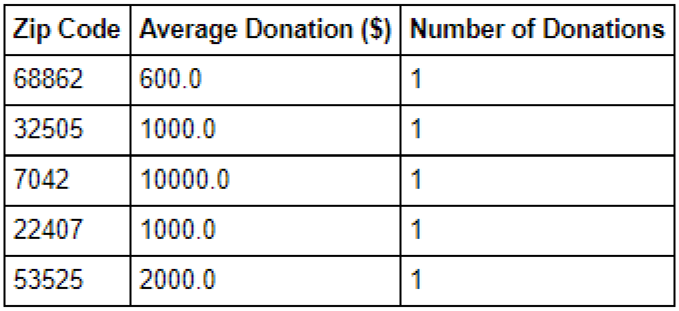
Zip Code	Average Donation ($)	Number of Donations
65049	5402615.0	1
33067	2480000.0	1
98407	2100000.0	1
67402	1647279.6666666667	3
60143	1290166.6666666667	3

Zip Code	Average Donation ($)	Number of Donations
58801	1.0	1
75701	1.0	1
68343	1.0	2
48030	1.0	3
22152	1.0	5

Figure 6. Left: Top 5 highest averaging zip codes. Right: Top 5 lowest averaging zip codes.



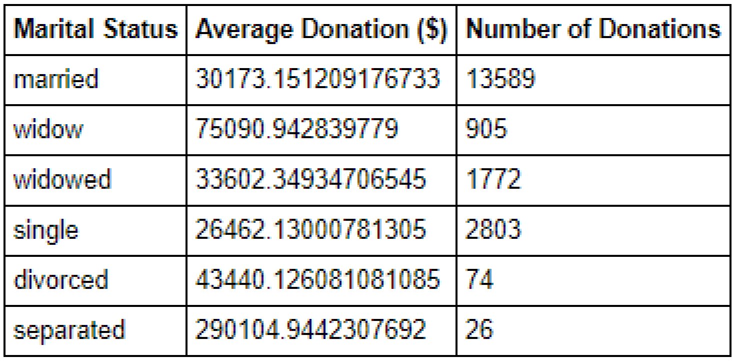
Zip Code	Average Donation ($)	Number of Donations
nan	96984.5844559586	386
32765	21659.160483870968	217
46514	2462.8750632911324	158
56031	26657.207337662316	154
68434	35225.86749999999	152



Zip Code	Average Donation ($)	Number of Donations
68862	600.0	1
32505	1000.0	1
7042	10000.0	1
22407	1000.0	1
53525	2000.0	1

Figure 7. Top 4 highest zip codes with greatest number of donations (excluding nan). Right: Top 5 zip codes with least number of donations.

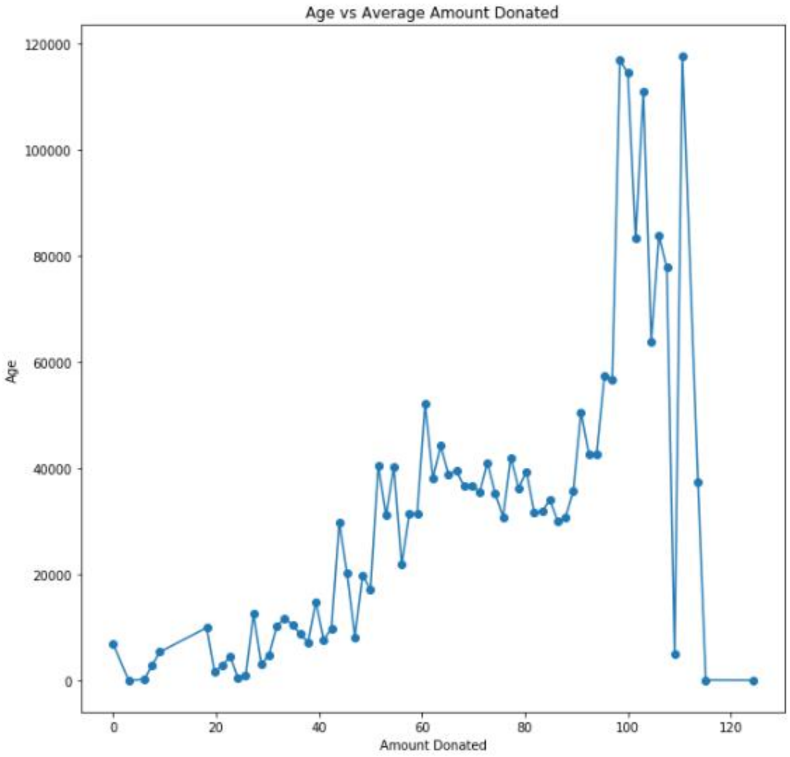
With regards to the marital status of the donor, we wanted to see the average of each category. Interestingly, widowed individuals donated the most on average, whereas single individuals donated the least on average.



Marital Status	Average Donation ($)	Number of Donations
married	30173.151209176733	13589
widow	75090.942839779	905
widowed	33602.34934706545	1772
single	26462.13000781305	2803
divorced	43440.126081081085	74
separated	290104.9442307692	26

Figure 8. Average donation of different marital statuses.

Age was also an interesting independent variable. It's intuitive that the older an individual is, the more money they have earned, and the more likely they are to donate. However, we did not expect to see a trend in which individuals who were close to or over a hundred years old tended to donate the most. Ages climbed as high as 120 years old. As it is unlikely that there exists so many centenarians, we figured the majority of people who are donating at such an advanced age are donating the rest of their worth to the Church after death. A quick cross-reference with the "deceased" column of our dataset confirmed these suspicions.



Finally, we analyzed the average and total donation amount from year to year, as well as the total number of donations each year. We observed that the trend for average donations stayed comparatively stable- there would be years where the average donation would peak, and years where the average donation would drop, but the trend line stays relatively flat. However, the trend for total donation amount as well as the total number of donations is a considerably sharp positive slope, although the data points seems to flatten in the recent few years. Finally, it is interesting to point out that just before 2010, there is a global peak in the total number of donations. While a peak also existed in the total donation amount, its value was far from close to the maximum peak, and the local peak in the average donation plot just prior to 2010 is actually underneath the average trend line. Unmistakably, many factors play into the donation amount, and there are complex interactions at play.

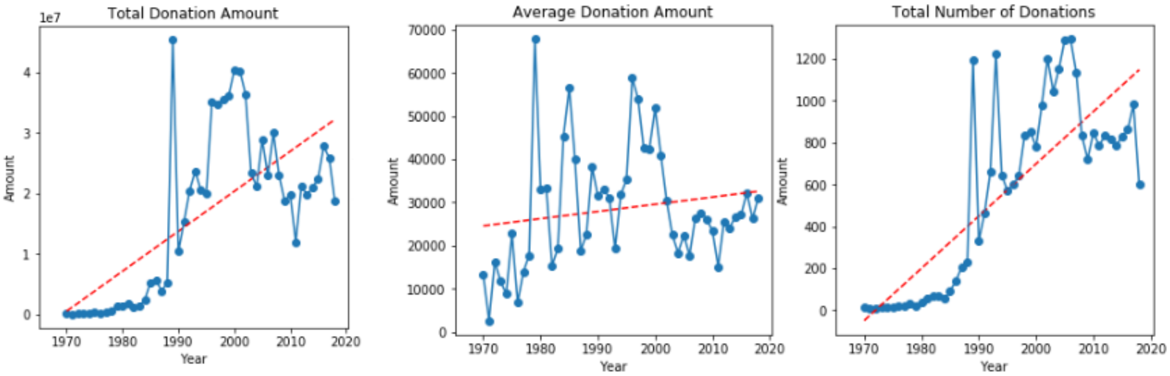


Figure 9. Figures derived from donations as a function of time

Now, we can begin to develop our models for our data, which we thoroughly understand. The first model we develop is a linear regression model. Our first linear regression model was simple- we used the year and the number of donations for that year as features, and we also used the total value of donations for that year as our label. We scaled and whitened our features, while placing slightly more weight on the number of people who donated to generate our model. While we expected an extremely poor score for this model, we cross validating our model 100 times and found a mean coefficient of determination (R^2 score) of 0.664- much better than what we had originally anticipated. The standard deviation of scores however, was larger than ideal, at 0.24.

We began considering what we could do to improve our linear regression model. We tried to break down our independent variables find a regression for each independent variable. For example, we found that on average, each age group donates different amounts, and that as years progress, the number of donors in each age group changes. Perhaps there was a linear pattern in how the number of individuals in each age group changes versus time. Following this idea, we performed 10 separate regressions on each of the 10 age groups, using a method that similarly follows the one we used when finding frequency of donation amounts. We divided the age distribution into buckets of even ranges, and measured how the count of each bucket changed over time. We performed a linear regression on each bucket, and used this data to predict the number of individuals donating from that particular age group for each year. Finally, we performed a linear combination of each of these age groups for any particular year, in which we multiplied the expected number of donations within that age group with the average amount per donation in that age group, and summed up the values over all bins. We saw a slight improvement in performance at an R^2 score of 0.691 and a standard deviation of 0.30.

However, developing a model from aggregating linear regression and linear combination only points us down into a deeper rabbit hole. For example, the average donation of each age group also changes from year to year- which is a pattern we did not include and must also try to find a linear regression for if we are to account for how each variable changes. Simply put, there are too many variables and dependencies to perform linear regression thoroughly. Another method must be used in order to have accurate and consistent results.

As we have the framework for the project built at this point, developing models should be an easier task moving forward. We are currently working on generating visuals for evaluation of hypothesis for our linear regression/linear combination model. Furthermore, we are also currently developing our kNN model, but we are running into issues regarding the high dimensionality of our data. In order to overcome this, we are considering using information gain to pick out features that have the most influence on our data, and use only those features in our kNN model. Following that, we will be develping a neural network model, as well as a random forest model, after which we will compare results across our different models and evaluate each of their performances.