ANA 610 Homework #2

Task #1 (90 pts):

In examining the Donor dataset (s_pml_donor_hw_v2), you realize that there are no dates. Yet you know that you will be asked to generate reports based on the following dates: date when individual was entered into the file ("origination date"), date of an individual's first gift ("first gift date"; and date of an individual's last gift ("last gift date").

1. Using SAS, create these fields on the dataset, giving them a format of MM/DD/YYYY (HINT – assume you are conducting this analysis on August 1, 1998 and each month has, on average, 30.4 days)).
   a. What is the date of the last file entry?
   03/02/1998

   b. What is the date of the first gift?
   12/10/1976

   c. What is the date of the last gift?
   04/01/1998

   d. What is median length of time (in months) between the first and last gift?
   48 months

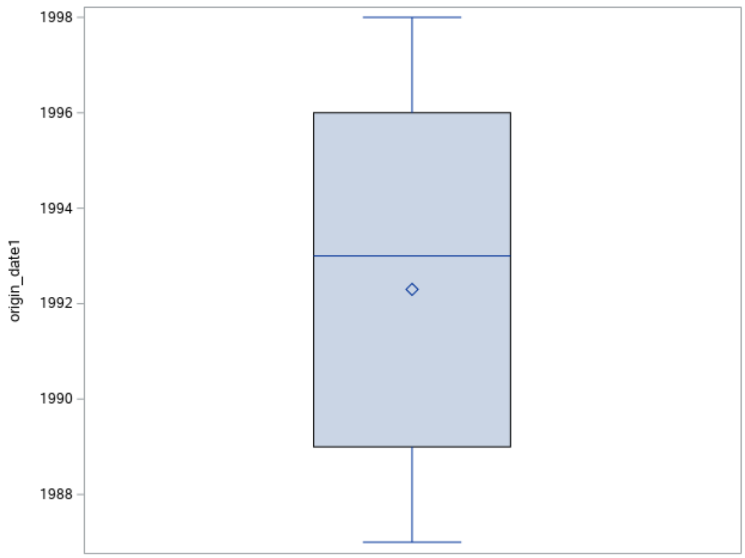
2. Using SAS, create 3 new fields, showing the YEAR (in YYYY format) for last file entry, first gift and last gift.
   a. In which year were the fewest number of individuals added to the file?
   1998

   b. Which year had the lowest average LAST_GIFT_AMT?
   1997

   c. Over the years, what has been the trend in the number of people donating ~~(use LAST_GIFT_AMT)~~?
   Over the years, there is a downward trend in the number of people donating. There was an increase between 1996 (7364 people) and 1997 (11433 people), but a significant decrease between 1997 (11433 people) and 1998 (575 people).

   d. In which STATE and YEAR was the average LAST_GIFT_AMT $17.963?
   Hawaii (HI), 1996

3. Using PROC SGPLOT,
   a. create and show the boxplot for the date when an individual was first entered into the file (origination date)



b. From this boxplot, in years, what is the IQR?
1989-1996 (7 years)

c. From this boxplot, what is the approximate median origination date in years?
1993

d. From this boxplot, what can you conclude about the skewness of the distribution of origination date?
The boxplot of the origination date shows that it is slightly negatively skewed.

4. Using PROC UNIVARIATE,
   a. Calculate the mean and median origination date in years from the "Basic Statistical Measures" output table. (Hint: use 365.25 to convert days into years.)
   Mean: 32.471 years (1992)
   Median: 33.171 years (1993)

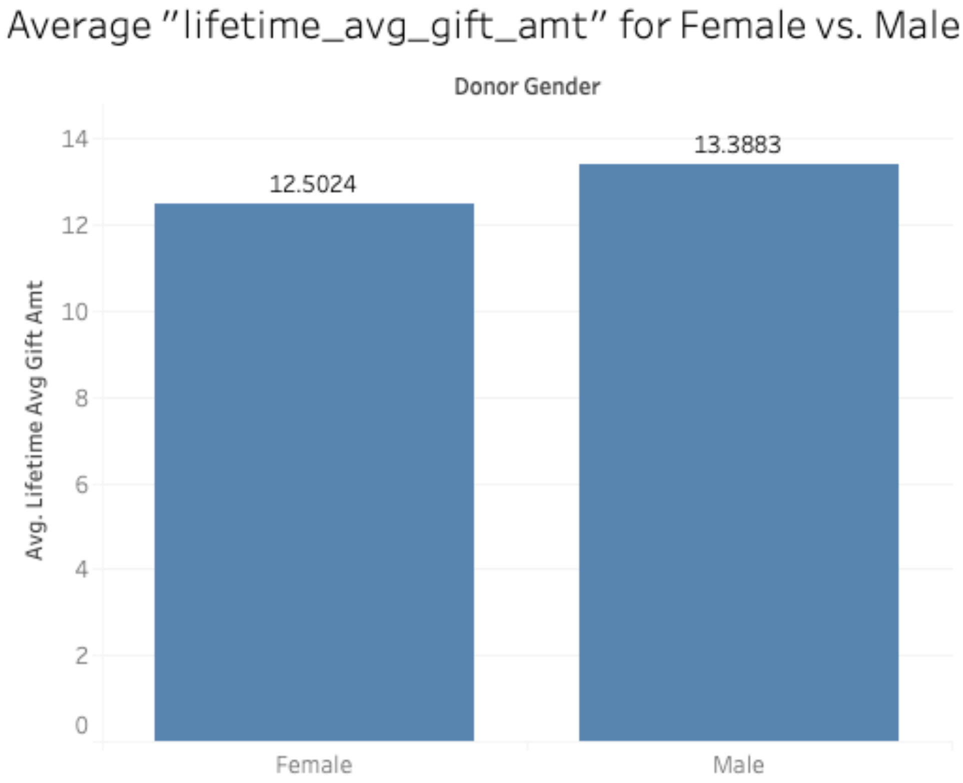
   b. Calculate the IQR from the "Quantiles" output table.
   6.99 years (2553.6 days)

   c. Do your calculated median and IQR match what you found above in #3? Explain
   Yes, it was very close. In Question 3, my IQR was 7 years, and the IQR I calculated is 6.99 years.

   d. Does the Skewness number produced by PROC UNIVARIATE match your conclusion on skewness from the boxplot? Explain
   Yes, it matches my conclusion from the boxplot and it is slightly negatively skewed (-0.23).

5. Using Tableau, create a bar chart showing the average LIFETIME_AVG_GIFT_AMT for female vs. male donors. Which gender has the higher average LIFETIME_AVG_GIFT_AMT? Label your bar charts to show these average values.

Males have a higher average "LIFETIME_AVG_GIFT_AMT".



Continuing our analysis of the Donor dataset, refer to "Data Dictionary – Donor.pdf" for data field definitions and your data audit performed in HW #1.

Task #2 (135 pts):

1. DONOR_AGE probably is an important variable in explaining whether an individual donated. However, it appears to have many missing values.
   a. Should these records be removed from the modeling dataset? Why or why not?
   No, we should not remove these records. DONOR_AGE has 4795 missing values, which is 32.89% of the values. Since this is less than 40%, it would be better to do a complete case analysis (if it was >40% it would be better to delete variable).

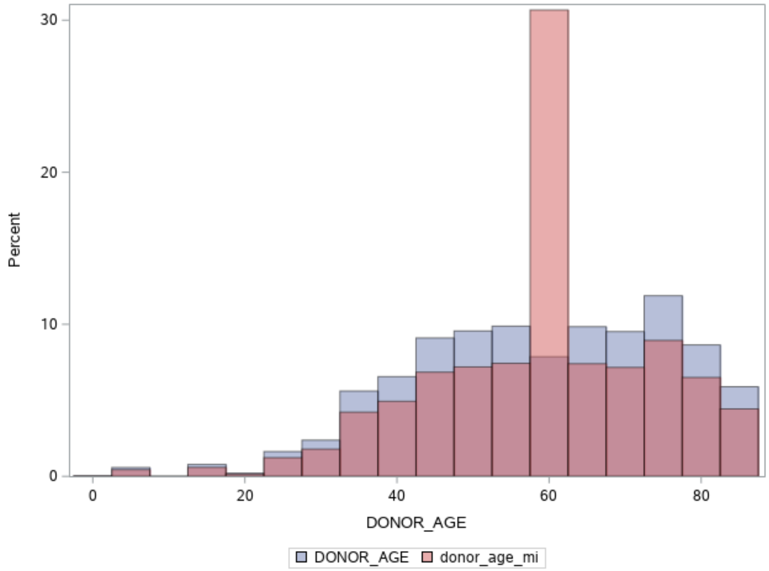
   b. Using SAS, perform a median-value imputation, a median-value segmentation imputation using 12 groups of MEDIAN_HOUSHOLD_INCOME, a hot-deck imputation using as additional variables MEDIAN_HOUSEHOLD_INCOME, MEDIAN_HOME_VALUE and MONTHS_SINCE_ORIGIN, and a stochastic regression imputation using the same additional variables (as used for hot-deck) as regressors.
      i. Report the findings of your imputations in the following table (report out to 3 decimal places):
      ii.

Technique	Seed value	Mean	Median	Mode	STD

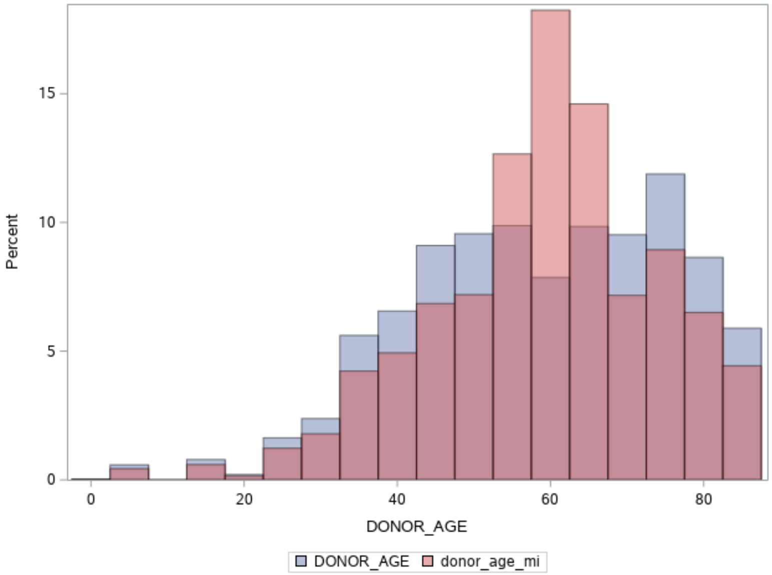
Observed values	N/A	58.919	60.000	67.000	16.669
Median-value	N/A	59.186	60.000	60.000	14.467
Median-value segmentation	N/A	59.248	61.000	63.000	14.526
Hot-deck	12345	58.918	60.000	67.000	16.643
Stochastic regression	12345	58.722	59.000	67.000	16.640

iii. For each imputation, show the “overlaid histogram” produced by PROC SGPLOT (imputed variable vs. observed variable). Make sure that the variable labels used by SGPLOT clearly differentiate between the observed and imputed variables. Use binwidth = 5 and transparency = .5.

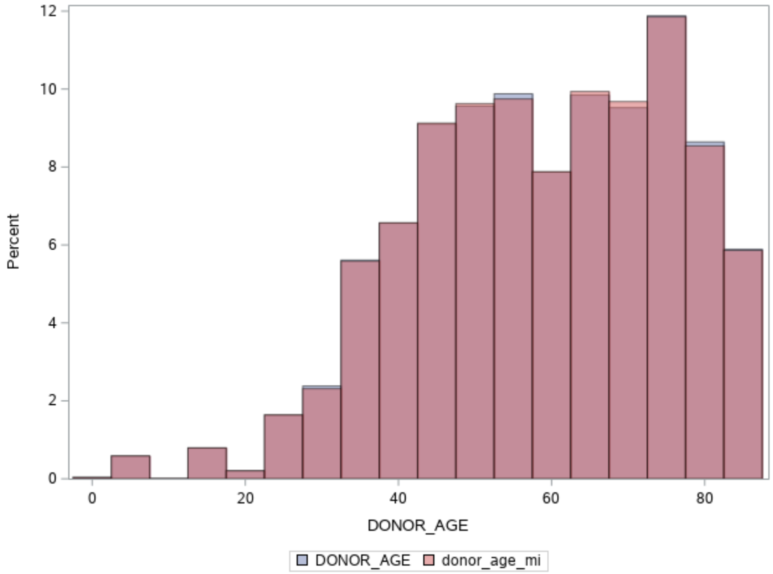
Median-value



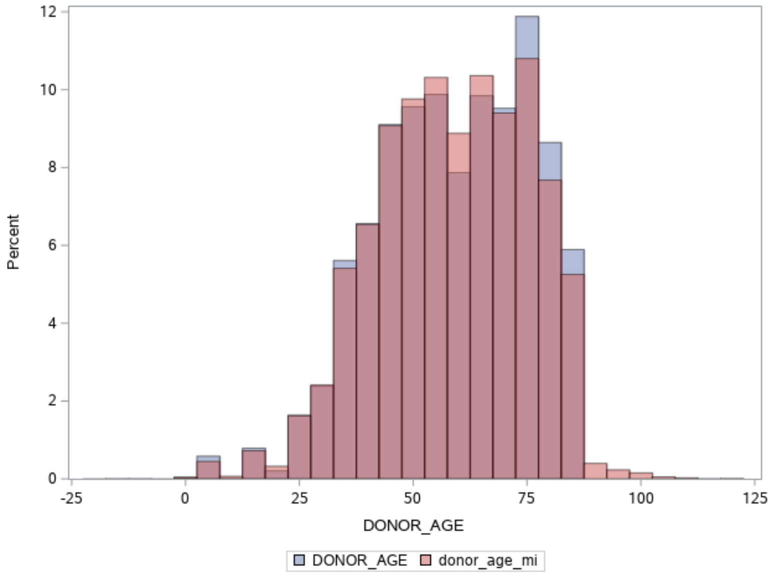
Median-value segmentation



Hot-deck



Stochastic regression



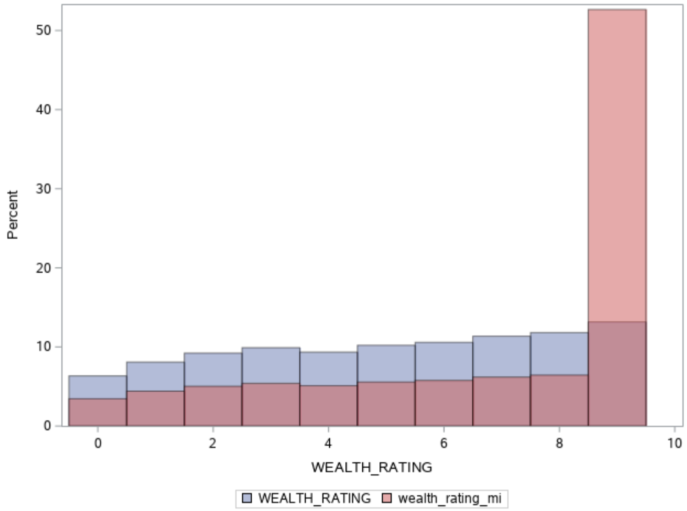
c. Based on the 4 imputations, which one do you recommend be used? Explain.
I would recommend using Hot-deck because the mean and variance (mean=58.91 std=16.64) is closest to the observed values (mean=58.91 std=16.66).

d. What should you always do when you impute missing values regardless of the technique? Explain.
When we impute, regardless of technique, we need to identify which technique yields close to the observed value mean and variance (standard deviation).

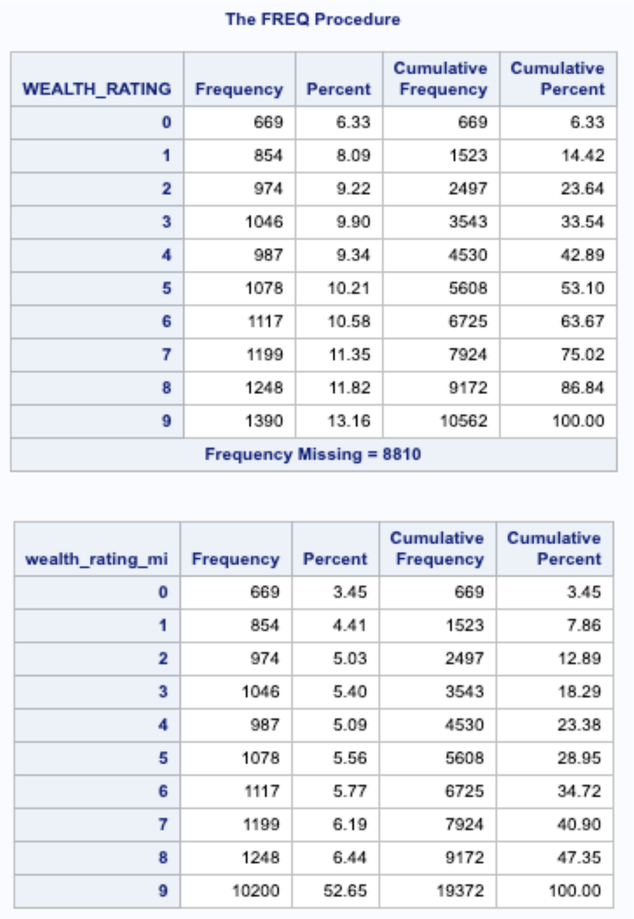
2. Imputing missing values for categorical variables, such as WEALTH_RATING, is not necessarily as straight forward as for continuous variables.

a. Using SAS, replace the missing values in WEALTH_RATING with the mode value and show the resulting:

   i. "overlaid histogram" produced by PROC SGPLOT (imputed variable vs. observed variable). Make sure that the variable labels used by SGPLOT clearly differentiate between the observed and imputed variables. Use binwidth = 1 and transparency = .5.



   ii. frequency of categories using PROC FREQ, both observed and imputes values.



The FREQ Procedure

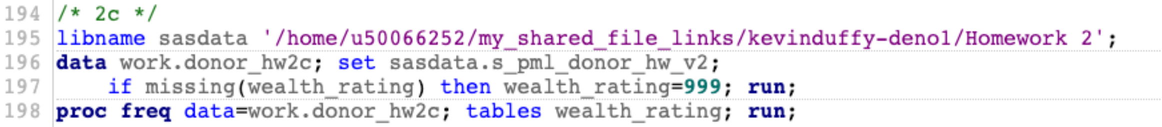
WEALTH_RATING	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	669	6.33	669	6.33
1	854	8.09	1523	14.42
2	974	9.22	2497	23.64
3	1046	9.90	3543	33.54
4	987	9.34	4530	42.89
5	1078	10.21	5608	53.10
6	1117	10.58	6725	63.67
7	1199	11.35	7924	75.02
8	1248	11.82	9172	86.84
9	1390	13.16	10562	100.00
Frequency Missing = 8810				

wealth_rating_mi	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	669	3.45	669	3.45
1	854	4.41	1523	7.86
2	974	5.03	2497	12.89
3	1046	5.40	3543	18.29
4	987	5.09	4530	23.38
5	1078	5.56	5608	28.95
6	1117	5.77	6725	34.72
7	1199	6.19	7924	40.90
8	1248	6.44	9172	47.35
9	10200	52.65	19372	100.00

b. Based on your SGPLOT histogram, would you recommend mode-value imputation for WEALTH_RATING? Explain.
No, I would not recommend mode-value imputation because my histogram shows a lot of variance between the observed values and the imputation values. The histogram shows that a wealth_rating of 9 shows the most variance, which is a strong indicator that mode-value imputation may not be the best choice.

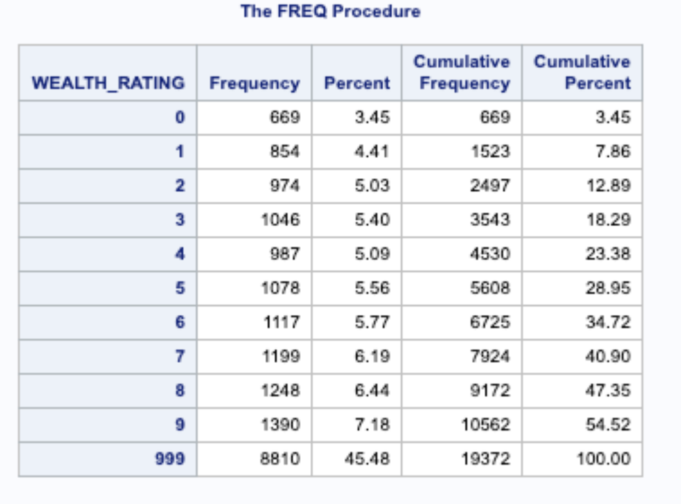
c. Show SAS code for a simple alternative for dealing with categorical variable missing values.
Note: In my code, I identified missing values as "999". In my output (proc freq table), the wealth_rating for "999" are all the missing values.

Code:



```
/* 2c */
libname sasdata '/home/u50066252/my_shared_file_links/kevinduffy-deno1/Homework 2';
data work.donor_hw2c; set sasdata.s_pml_donor_hw_v2;
    if missing(wealth_rating) then wealth_rating=999; run;
proc freq data=work.donor_hw2c; tables wealth_rating; run;
```

Output:



The FREQ Procedure

WEALTH_RATING	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	669	3.45	669	3.45
1	854	4.41	1523	7.86
2	974	5.03	2497	12.89
3	1046	5.40	3543	18.29
4	987	5.09	4530	23.38
5	1078	5.56	5608	28.95
6	1117	5.77	6725	34.72
7	1199	6.19	7924	40.90
8	1248	6.44	9172	47.35
9	1390	7.18	10562	54.52
999	8810	45.48	19372	100.00

NOTE – WEALTH_RATING is already machine numeric. So, you do NOT need to create a numeric version as shown in the lecture slides and can proceed directly to the mode calculation and imputation stages.

Extra Credit (20 pts)

Variables with extreme or non-normal distributions make hypothesis testing difficult and can adversely affect model fit, depending on the algorithm employed.

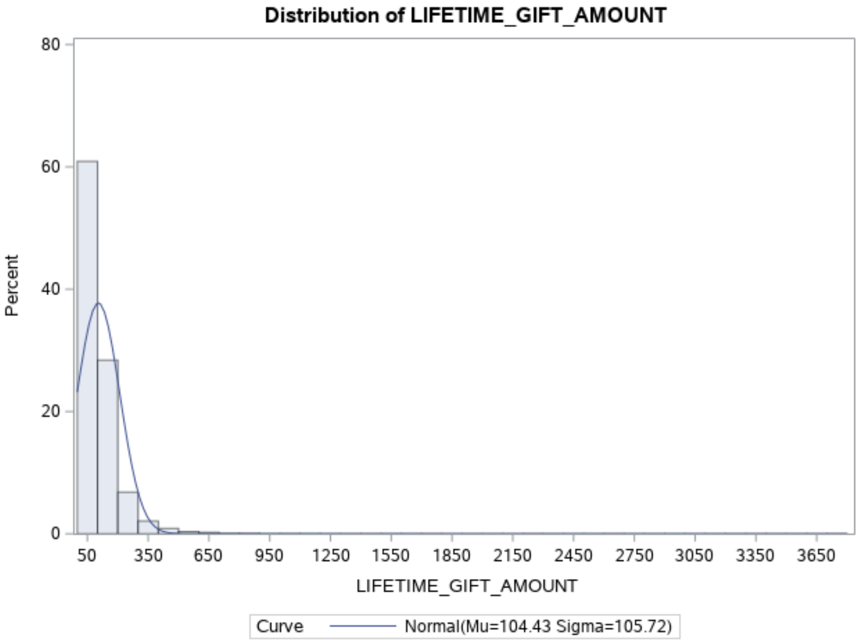
1. Is the distribution of the variable LIFETIME_GIFT_AMOUNT extreme? Explain. Upon what statistic are you basing your analysis?
   Yes, the distribution of LIFETIME_GIFT_AMOUNT is extreme. The skewness statistic for this distribution is 6.59, which is over 5. If the abs(skewness) > 5, then the distribution is extreme.

2. If it is extreme, identify the single transformation which yields the lowest (absolute value) skewness statistic (try square, square root, inverse, inverse of square root, inverse of square and log). Show the before and after histogram from PROC UNIVARIATE as well as the before and after skewness statistic.

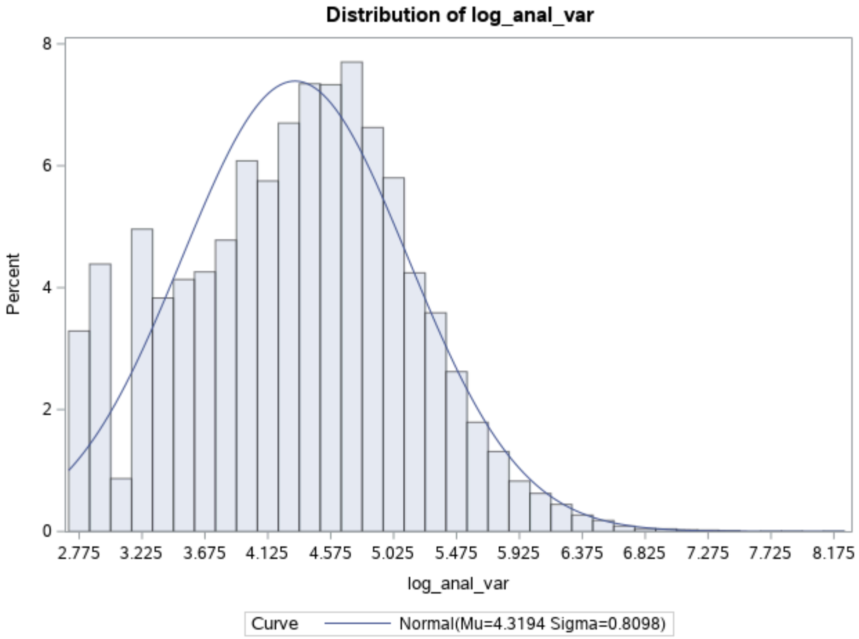
Transformation	Skewness
No transformation (original)	6.593
Square	65.405
Square Root	1.494
Inverse	1.502
Inverse of square root	0.811
Inverse of square	2.679
Log	0.041

Transformation using Log yields the lowest (absolute value) skewness statistic (0.04).

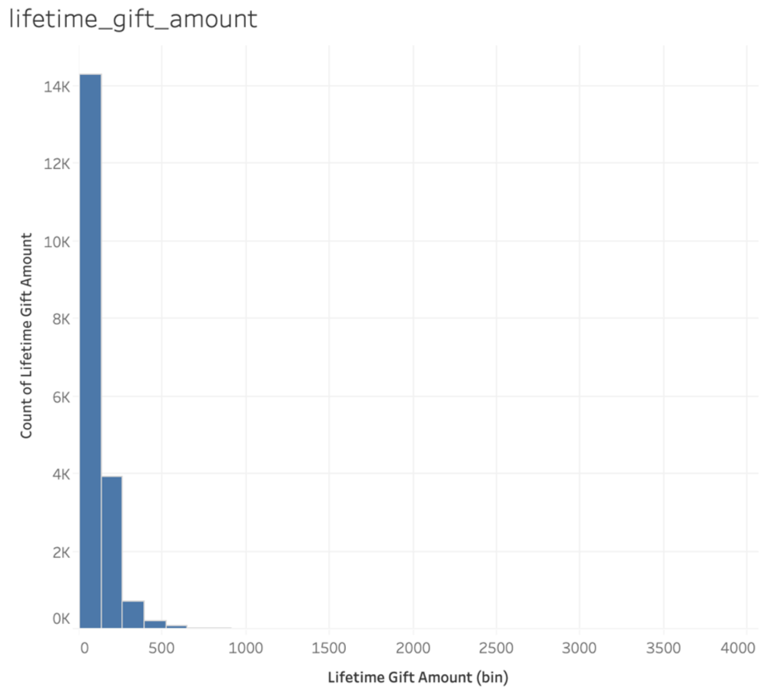
Before Histogram (Skewness=6.59):



After Histogram (Skewness=0.04):

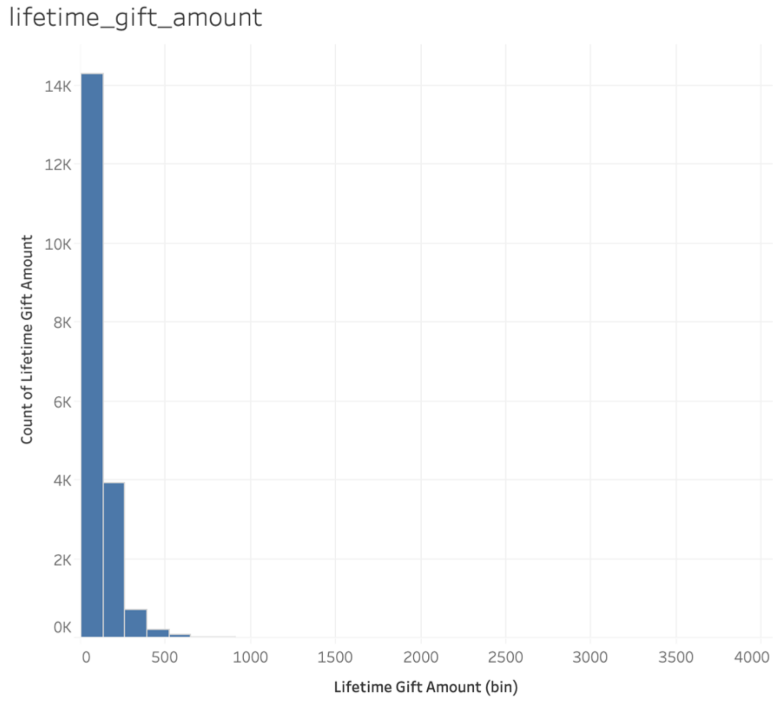


3. Using Tableau and Tableau's default bin size:
    a. create and show the histogram for LIFETIME_GIFT_AMOUNT;

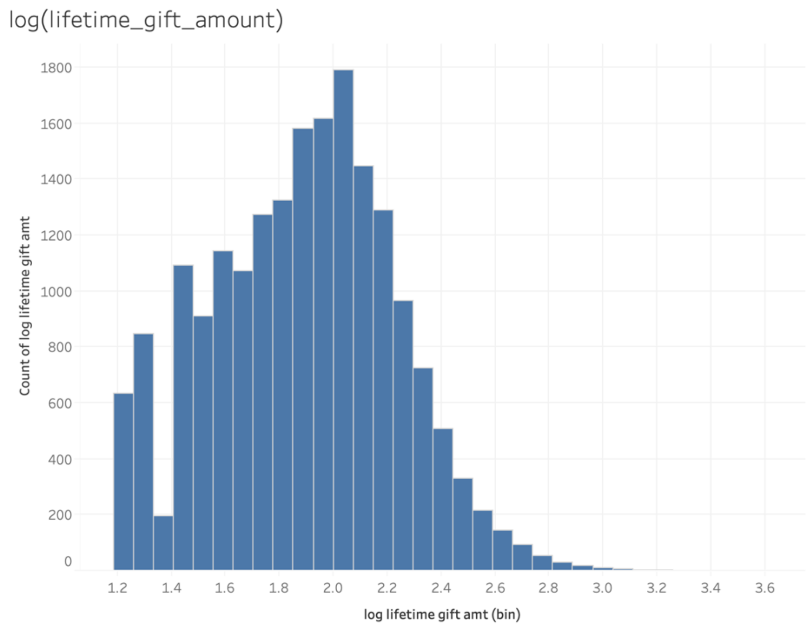


b. create the appropriate transformation used above. Show the histogram for this transformed field next to the untransformed histogram.

Untransformed (Before):



Transformed (Log):



Homework deliverables:

- Tasks 1 – 2 plus extra credit:
  - separate Word doc with your analysis and discussion, including all tables and charts
  - SAS program (as a .sas file) with all code used for Tasks 1 – 2 plus extra credit