**ANA 610 Homework #2**

**Please read directions carefully and refer to the “grading rubric” for point allocation.**

**Task #1 (68 pts):**

In examining the 19,000+ row Donor datafile (s\_pml\_donor\_hw\_v3) **and the associated data dictionary**, 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 an individual was entered into the file (“ENTRY\_DATE”),
* date of an individual’s first gift (“FIRST\_GIFT\_DATE”), and
* date of an individual’s last gift (“LAST\_GIFT\_DATE”).

**Using SAS:**

1. Using the fields MONTHS\_SINCE\_ORIGIN, MONTHS\_SINCE\_FIRST\_GIFT, and MONTHS\_SINCE\_LAST\_GIFT and the SAS INTNX() function, create these date fields on the datafile, giving them a format of MM/DD/YYYY. Assume you are conducting this analysis on August 1, 1998.
   1. What is the date of the most recent file entry? 03/01/1998
   2. What is the date of the first gift? 12/01/1976

(The date of the most recent first gift: 05/01/1997)

* 1. What is the date of the last gift? 04/01/1998
  2. What is the median length of time (in months) between the first and last gift? 48

1. Using PROC UNIVARIATE with default settings,
   1. Show the histograms for ENTRY\_DATE and FIRST\_GIFT\_DATE **with the year displayed on the x-axis.**

1988

1988

* 1. Based on these two histograms, what is odd about ENTRY\_DATE and FIRST\_GIFT\_DATE? Explain. The earliest entry date is March 1987. There are 30 FIRST\_GIFTS that occur before that date, from Dec 1976 up to Feb 1987. This data integrity issue needs to be addressed and a decision needs to be made on what to do with those observations.

1988

1. Using the SAS YEAR() function, create 3 new fields, showing the YEAR (in YYYY format): ENTRY\_DATE\_YEAR, FIRST\_GIFT\_DATE\_YEAR, and LAST\_GIFT\_DATE\_YEAR.
   1. How many individuals were added to the file in 1998? Support your answer by showing an SAS output table. 7

| **ENTRY\_DATE\_YEAR** | **Frequency** | **Cumulative Frequency** |
| --- | --- | --- |
| **1998** | 7 | 7 |

* 1. Which year had the lowest mean LAST\_GIFT\_AMT? Support your answer by showing an SAS output table using LAST\_GIFT\_DATE\_YEAR . 1997

| **Analysis Variable : LAST\_GIFT\_AMT** | | |
| --- | --- | --- |
| **LAST\_GIFT\_DATE\_YEAR** | **N Obs** | **Mean** |
| 1996 | 5718 | 18.430 |
| 1997 | 12988 | 15.711 |
| 1998 | 666 | 17.770 |

* 1. In CLUSTER\_CODE = 9 and LAST\_GIFT\_DATE\_YEAR = 1997, what was the **mean** LAST\_GIFT\_AMT? Support your answer by showing an SAS output table again using LAST\_GIFT\_DATE\_YEAR . 14.768

| **Analysis Variable : LAST\_GIFT\_AMT** |
| --- |
| **Mean** |
| 14.768 |

**Using R:**

1. **Show here**
   1. create the variables ENTRY\_DATE, FIRST\_GIFT\_DATE, LAST\_GIFT\_DATE
   2. output from summary.Date() with min, med, mean and max values for above.
2. **Show here**
   1. R code for how you would create the variable LAST\_GIFT\_DATE\_YEAR
   2. output showing mean LAST\_GIFT\_AMT by LAST\_GIFT\_YEAR (match 3(b)).

**Task #2 (133 pts):**

1. The AnyState Veterans of Foreign Wars solicits donations from adults aged 18+ years. DONOR\_AGE may be an important variable in explaining whether an individual donated and, if so, how much they donated.
   1. Using a **table** to summarize your findings, perform a quick data integrity check (using R or SAS) on DONOR\_AGE and identify issues which may exist with respect to this variable. Check and report briefly (**including a brief note per item if there is any concern**) on missing values, odd values, extreme values on the low and high end, distribution skewness, value consistency and anything else you think is important.

| **Analysis Variable : DONOR\_AGE** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **N** | **N Miss** | **Minimum** | **Maximum** | **Mean** | **Std Dev** |
| 14577 | 4795 | 0 | 87 | 58.919 | 16.669 |

4795 missing values is a large portion of the overall observations.

206 values for minor ages 0-17, this range does not fit the guidance and needs addressing.

8 values are identified as outliers 18<values<21, seems okay to keep, confirm.

Q1=47, Q2=60, Q3=73, 1.5\*26=39, outliers: 21>values>99, est. ~ close to obs.

| **Quantiles for Normal Distribution** | | |
| --- | --- | --- |
| **Percent** | **Quantile** | |
| **Observed** | **Estimated** |
| **5.0** | 31 | 31.500 |
| **25.0** | 47 | 47.676 |
| **50.0** | 60 | 58.919 |
| **75.0** | 73 | 70.162 |
| **95.0** | 83 | 86.338 |

* 1. One integrity issue you just uncovered is missing values. Should these records (rows) with missing values for DONOR\_AGE be removed from the modeling datafile? **Explain fully** why, or why not? No, do not remove --

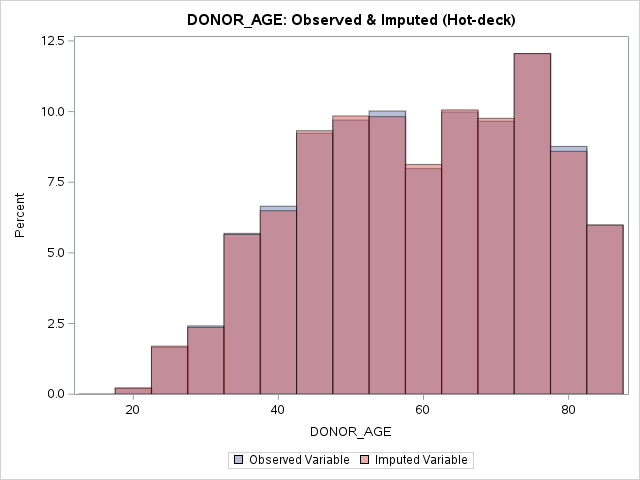
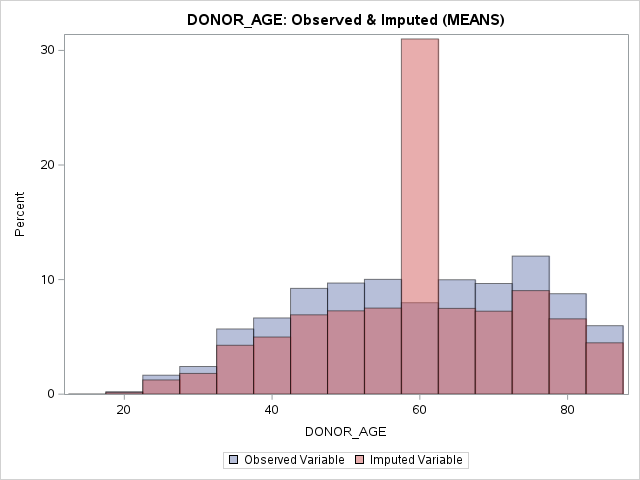
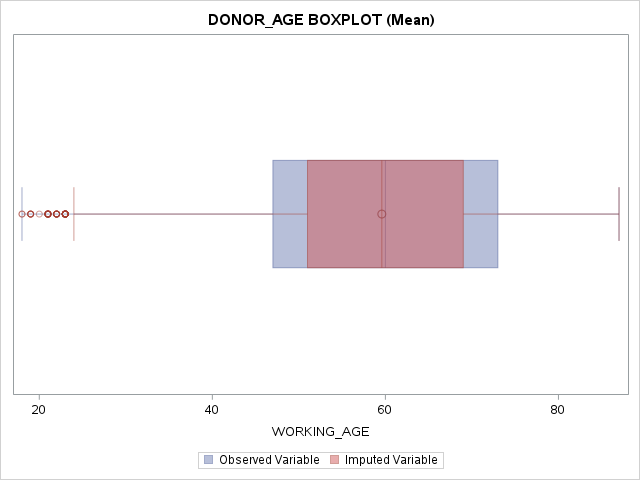
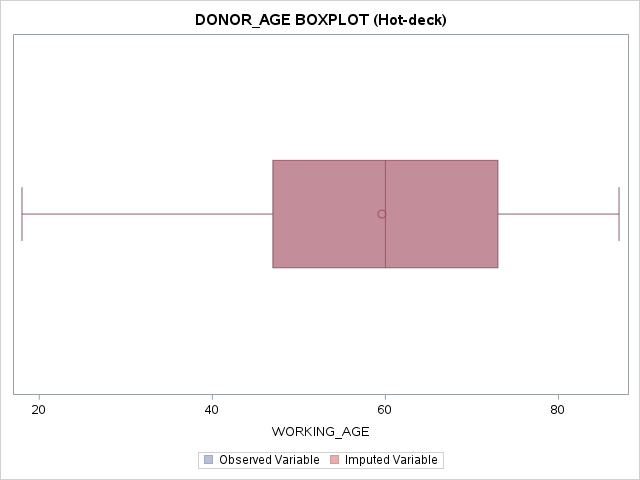
Missing values represent 25% of the records, a significant portion, removing them could lead to a loss of information and impact the model’s performance. We will have to analyze how much of an impact age has on donations: Is age a key variable that is crucial to the predictive model, then missing values must be addressed. If its exclusion has minimal impact on the model, removing missing values may be okay. Are missing values random (MCAR, MAR, or MNAR)?

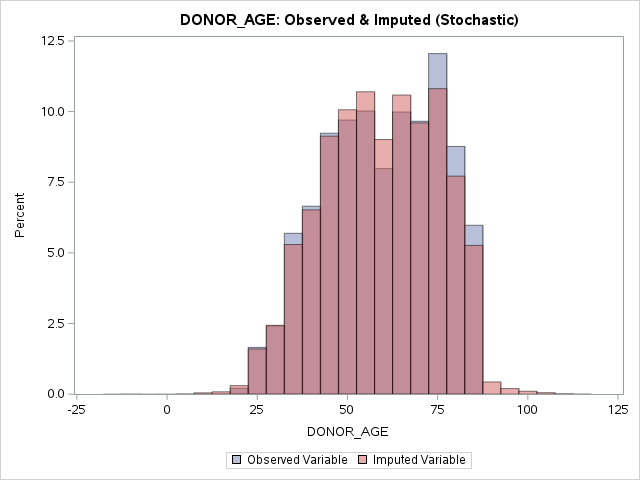
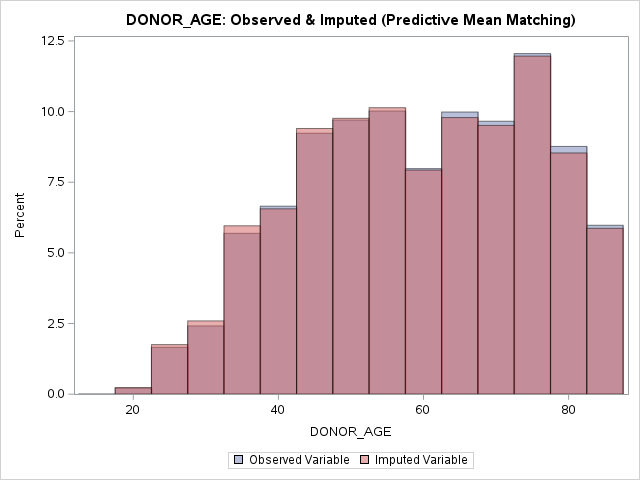
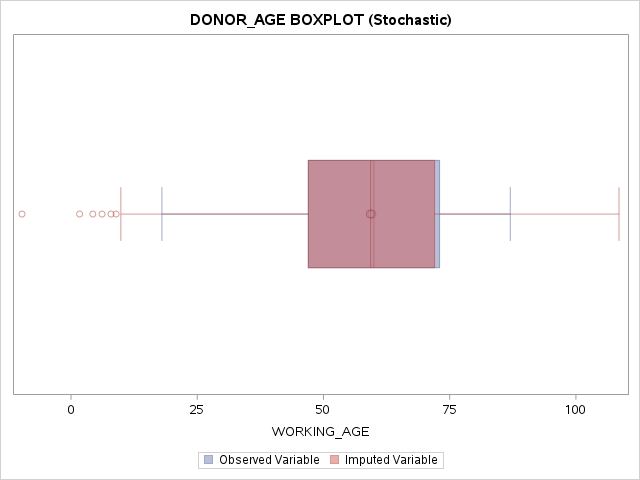
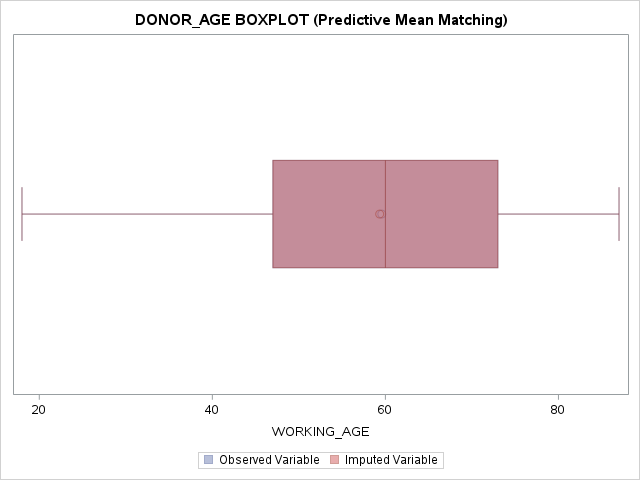
Depending of their randomness, removing missing values may or may not introduce bias. Use SAS to investigate whether the missing values are randomly distributed or have a relationship to other variables. We may even want to run two models -- one with missing values imputed and another with rows dropped -- and compare results to see if the removal impacts our conclusions.

1. **Using SAS**, impute the missing values of DONOR\_AGE using **mean unconditional** and the **conditional** approaches **hot-deck**, **stochastic** **regression**, and **predictive mean matching (PMM)**. **Your analysis will make use of the following**:

* Upon guidance from AnyState, the modeling datafile should be limited to those records with either missing values for DONOR\_AGE or values for DONOR\_AGE >= 18.
* The control variables for the conditional approaches are months\_since\_origin, in\_house, pep\_star, lifetime\_card\_prom, lifetime\_prom, and months\_since\_first\_gift.
* Months\_since\_origin will be used for the correlation metric.
  1. Report the findings of your imputations in the following table (report out to 3 decimal places):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Imputed Values** | | | |
| **Statistic** | **Observed Values** | **Mean** | **Hot-Deck** | **Stochastic Regression** | **PMM** |
| **Seed value =>** | | | **12345** | **12345** | **12345** |
| **N** | 14,371 | 19,166 | 19,166 | 19,166 | 19,166 |
| **Min** | 18 | 18 | 18 | -9.712 | 18 |
| **Mean** | 59.587 | 59.587 | 59.602 | 59.273 | 59.339 |
| **Median** | 60 | 59.587 | 60 | 59.347 | 60 |
| **Max** | 87 | 87 | 87 | 108.536 | 87 |
| **STD** | 15.808 | 13.688 | 15.810 | 15.821 | 15.853 |
| **Correlation** | 0.276 | 0.238 | 0.210 | 0.281 | 0.278 |

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



* 1. Based on the 4 imputations, which one do you recommend be used? **Explain fully**. Predictive Mean Matching:

The mean method preserves the central tendency well, but typically underestimates variance, reflected in a similar std dev as the observed data. This could limit variability.

Hot-deck imputation also preserves the central tendency well but adds more variability to the observed data, (higher std dev). The reduced correlation suggests a weaker relationship between imputed values and the observed data.

With a high correlation, stochastic regression also introduces extreme values (negative min and high max), and negative are not realistic for DONOR\_AGE. These extremes could distort the analysis if not addressed.

PMM is a good balance between preserving the central tendency and maintaining a strong correlation with the observed data. The variability is slightly increased, but this could have been a more realistic reflection of the missing data's variability.

* 1. What should you always do when you impute missing values regardless of the technique? **Explain fully.**

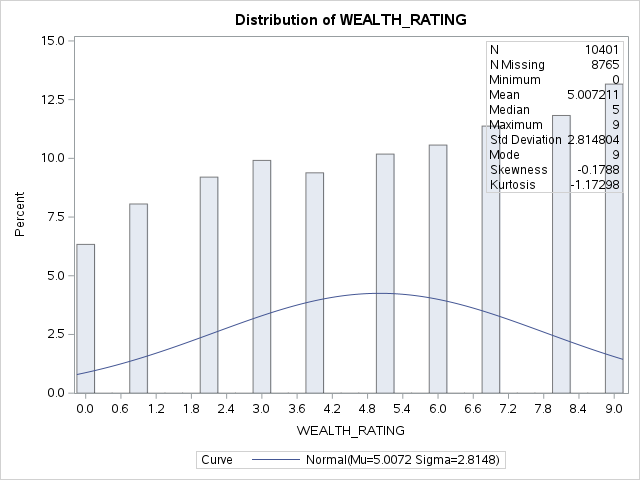
Create a missing indicator that tracks the original MVs as 0 or 1. Create a new imputed variable so you do not overwrite the original variable. Document your process to provide context so those seeing the data understand and can replicate without you.

1. Imputing missing values for categorical variables, such as WEALTH\_RATING, is not necessarily as straight forward as for continuous variables.
   1. Using a table, perform a quick data integrity check (using R or SAS) on WEALTH\_RATING and identify all issues with respect to this variable. Check and report briefly (**including a brief note per item if there is any concern**) on missing values, odd values, extreme values on the low and high end, distribution skewness, value consistency and anything else you think is important.

8810 MVs (46%), This is the only issue of concern, almost half of the observations.

IQR=4, 1.5\*4=6, no outliers, est. are approximate to obs. kurtosis -1.173=a bit flat

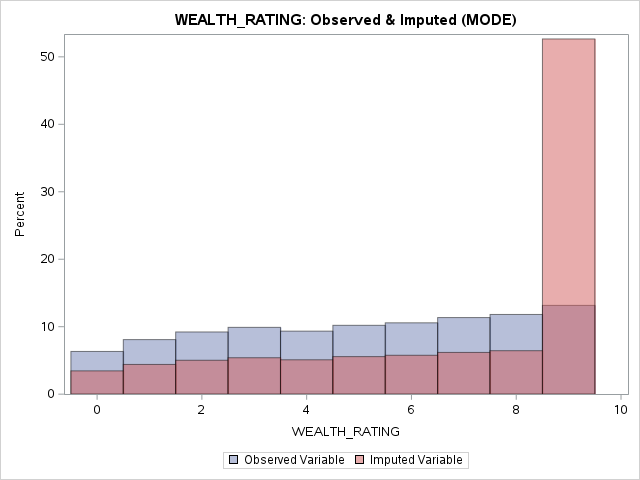
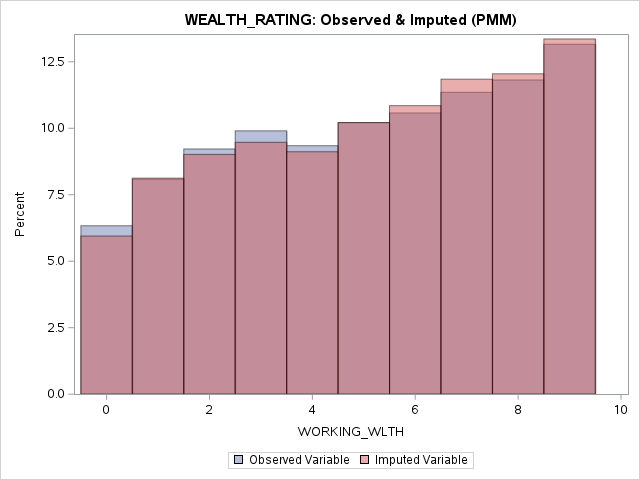
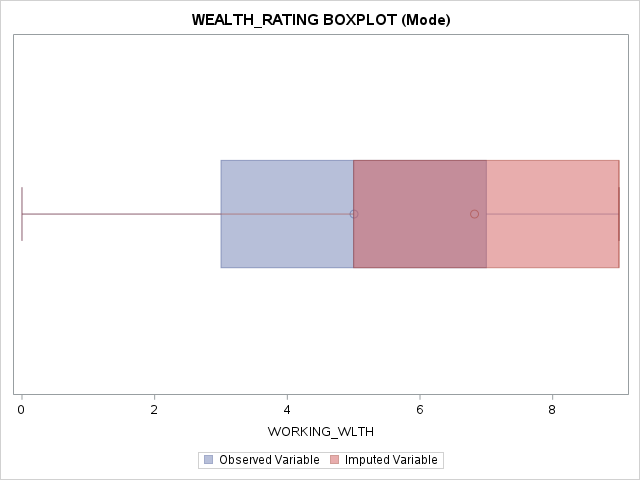
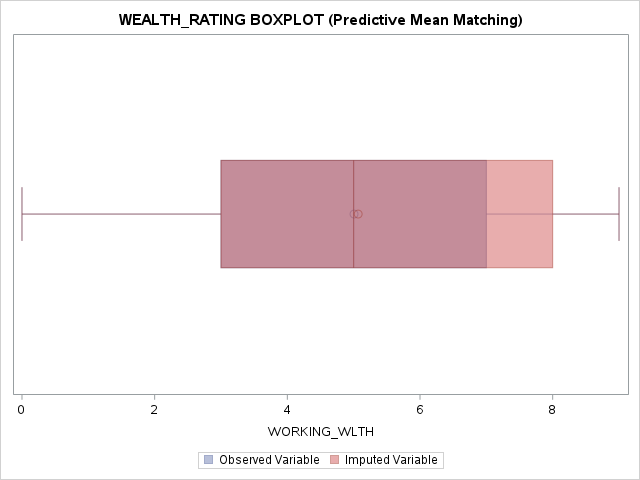
| Analysis Variable : WEALTH\_RATING | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **N** | **N Miss** | **Minimum** | **Mean** | **Median** | **Maximum** | **Std Dev** | **Mode** |
| 10,562 | 8810 | 0 | 5.005 | 5 | 9 | 2.815 | 9 |



| **Quantiles for Normal Distribution** | | |
| --- | --- | --- |
| **Percent** | **Quantile** | |
| **Observed** | **Estimated** |
| **5.0** | 0 | 0.375 |
| **25.0** | 3 | 3.106 |
| **50.0** | 5 | 5.005 |
| **75.0** | 7 | 6.904 |
| **95.0** | 9 | 9.636 |

* 1. **Using SAS**, impute the missing values of WEALTH\_RATING using the **mode unconditional** and **PMM conditional** approaches.
* The modeling datafile should include ALL 19,372 records (rows).
* The control variables for the PMM conditional approach are median\_home\_value, pep\_star, and per\_capita\_income.
* Per\_capita\_income will be used for the correlation metric.
  + 1. Report the findings of your imputations in the following table (report out to 3 decimal places):

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Imputed Values** | |
| **Statistic** | **Observed Values** | **Mode** | **PMM** |
| **Seed value =>** | | | **12345** |
| **N** | 10,562 | 19,372 | 19,372 |
| **Min** | 0 | 0 | ~0 |
| **Mean** | 5.005 | 6.822 | 5.068 |
| **Median** | 5 | 9 | 5 |
| **Mode** | 9 | 9 | 9 |
| **Max** | 9 | 9 | 9 |
| **STD** | 2.815 | 2.877 | 2.805 |
| **Correlation** | 0.521 | 0.270 | 0.505 |

* + 1. For each imputation, show here the “overlaid histogram” and “overlaid boxplot” 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** (histogram) and transparency = .5 (histogram and boxplot).
  1. Based on the 2 imputations, which one do you recommend be used? **Explain fully. PMM**

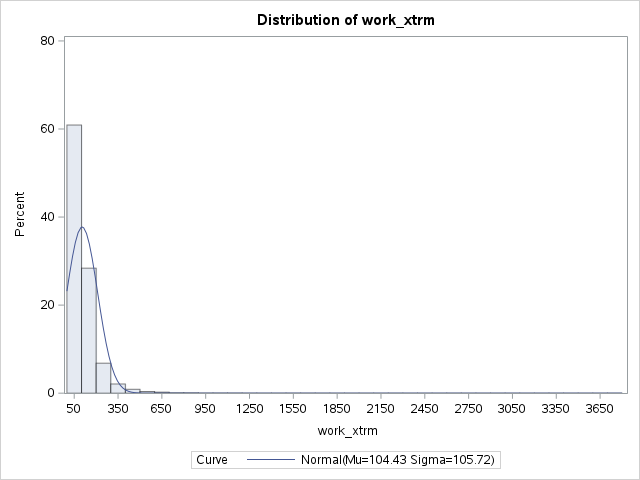
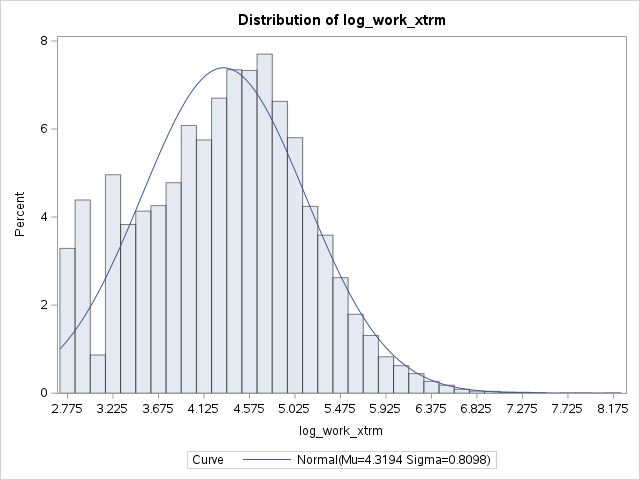
PMM is close to obs mean/med; it preserves the central tendency. Single Mode distorts mean and shifts med to 9. PMM is close to std dev of the obs data, preserving variability. Single Mode introduces slightly more variability. PMM keeps a correlation significantly closer to the obs than Single Mode: Better retaining the relationship between the variables in the dataset. Overall, PMM combines the best options by leveraging observed values to impute missing ones, ensuring the imputed values are realistic and consistent with the rest of the data. Mode imputation oversimplified the data by assigning the mode value to all missing values, which lead to overestimation and distortion of the data.

**Task #3 (19 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 fully**. Upon what statistic are you basing your analysis? Use the entire 19,372 row datafile.

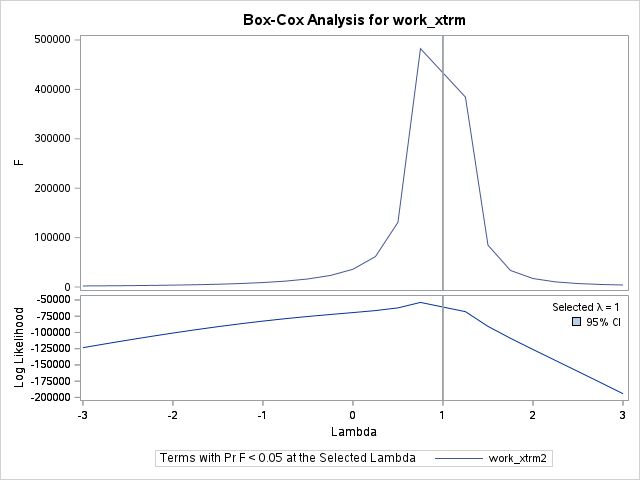
Comparing the mean (104) and median (79) shows that the distribution is extreme and highlights the skewness. This tells us that a majority of the values are low, to keep the median low, but there are extremely high values that increase the mean.

1. If it is extreme, manually using Tukey’s Ladder of Powers, identify the single transformation which yields the most normal distribution. Show the before and after histogram from PROC UNIVARIATE as well as the before and after skewness statistic.

| **Moments** | | | |
| --- | --- | --- | --- |
| **Skewness** | 6.594 | **Kurtosis** | 124.148 |

| **Moments** | | | |
| --- | --- | --- | --- |
| **Skewness** | 0.041 | **Kurtosis** | -0.331 |

1. Using a BoxCox transformation, can the skewness be further reduced? **Explain fully**, including in your answer: a. Output chart produced by PROC TRANSREG

b. The optimal lambda and

the new skewness value.

| **Obs** | **Box** | **λ** | **log** | **Inv\_sqrt** |
| --- | --- | --- | --- | --- |
| **skew** | 6.594 | 1 | 0.041 | 0.811 |

The high skewness and lambda = 1 indicate that even after trying the Box-Cox, the data still has significant skewness. There is some characteristic that limits the effectiveness of the Box-Cox in reducing skewness further. Inv\_sqrt looks perfect in the plot until you look at the x-axis

**Extra Credit (20 pts)**

1. **Using R**,
   1. **Show here** your R code for how you would use K-nearest neighbor (kNN) to impute the missing values of DONOR\_AGE using kNN() from the VIM package.
   2. Add a column to the table you filled in in Task #2 which reports the stats obtained from using kNN.
   3. Do these new findings change your recommendation as to which imputation method to use? **Explain fully**.
2. **Using SAS’s** PROC UNIVARIATE, **the ENTRY\_DATE variable you created in TASK #1**, and the output table “Basic Statistical Measures”:
3. By hand, using 365.25 to convert days into years, **calculate the year** of the median entry date out to two decimal places (i.e., XXXX.XX). Support your answer by showing the Basic Statistical Measures output table.
4. Using SAS, **calculate and show** the same median entry date ENTRY\_DT\_MEDIAN, but in MM/DD/YYYY format, using the INTNX() function and a new variable MEDIAN\_DAYS set equal to the median value shown in the Basic Statistical Measures output table. **Also show your code**.
5. **Explain fully** whether the values in (a) and (b) are consistent with each other.

**Homework deliverables:**

* Tasks 1 – 3 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 – 3 plus extra credit.