Final Project

STAT 7100 – Statistical Methods

For Dr. Kimberly Gardner

By Matthew Connor Armstrong

Due November 12th, 2019

Graduate College of Science and Mathematics

Kennesaw State University

Abstract

This paper is written with the purpose of fulfilling the final requirement for STAT-7100. An open-source dataset containing basic information related to meteorites and their respective landing sites will be utilized to answer two distinct research questions pertaining to the dataset, as well as an overarching research question for this data, to be discussed further within the body of the report. The statistical software suite SAS (Statistical Analysis System) and data visualization software Tableau will be implemented to complete this exercise.

*Keywords:* Statistics, Meteorites, ANOVA, Analytics, SAS

Final Project

This project will be primarily concerned with the dataset Meteorite\_Landings.csv found in NASA’s Open Data Portal at this URL: https://data.nasa.gov/Space-Science/Meteorite-Landings/gh4g-9sfh

This dataset has 9 columns of information which describe the meteorite physically, its geographic location, and the date it fell or was found. See Table 1 for a list of variables (ie columns) and their descriptions.

Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Variable Type** | **Description** | **Example(s)** |
| Name | Text | Name of Meteorite | Northwes, Dhofar 1 |
| ID | Numeric – Integer | ID of Meteorite | 50693, 1, 11927 |
| NameType | Text – Binary | Type of Name | Relict, Valid |
| RecClass | Text | Type of Meteorite | L6, H5, Howardit, CM2 |
| Mass\_g | Numeric – Continuous | Mass of Meteorite | 256.8, 951, 33.4 |
| Fall | Text – Binary | Fell or Found | Fell, Found |
| Date | Numeric – Integer | Year fell or found | 2003, 1900, 1901 |
| RecLat | Numeric – Continuous | Geographic Latitude | 33.45, -26.66667, 23.9 |
| RecLong | Numeric - Continuous | Geographic Longitude | 54.5875, 108.1, -90.221 |

There are many types of analysis which could be performed on this type of data. In the following sections this data will be analyzed using a One-way ANOVA and a Chi-square test of independence. These statistical methods will be utilized to determine whether the 4 most common instances of the variable “recclass” are correlated in some way to the variables “mass\_g” (a continuous variable) and “Fall” (a binary categorical variable).

**Part 1: Methods for More than Two Samples – One-Way ANOVA**

In this section, the masses for the 4 most commonly found/seen meteorites will be analyzed using a one-way ANOVA. This method allows one to determine whether a case can be made that one or more of the means “does not originate from the same population.” This type of analysis could be useful for determining whether the variance among a certain population of meteorites was due to chance or some other confounding factor. Specifically, this section will address whether the average masses of the 4 most common types of meteorites are sufficiently different to validate that at least one of the average masses of these types of meteorites vary to a 95% degree of confidence (0.05 significance level).

Null Hypothesis H0:

Avg mass (L6) = Avg mass (H5) = Avg mass (L5) = Avg mass (H6)

Alternate Hypothesis HA:

The average masses of L6, H5, L5, and H6 meteorites are not all equal.

Independent Variable: recclass, type of meteorite

Dependent Variable: mass\_g, mass of meteorite in grams

The data which did not meet the condition – recclass = L6, H5, L5, or H6 was removed from the dataset and the following code was ran in SAS:

**proc** **ANOVA** data=project.anova;

title One-way ANOVA Comparing the Masses of L6, H5, L5, and H6 Meteorites;

class recclass;

model mass\_g = recclass;

means recclass;

**run**;

Resulting in the following output:

**One-way ANOVA Comparing the Masses of L6, H5, L5, and H6 Meteorites.**

Table 2

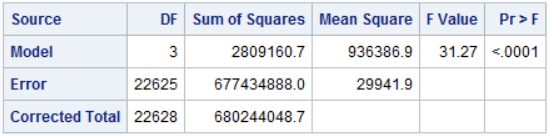


Table 3

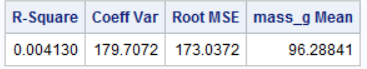


Table 4

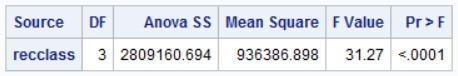


Figure 1

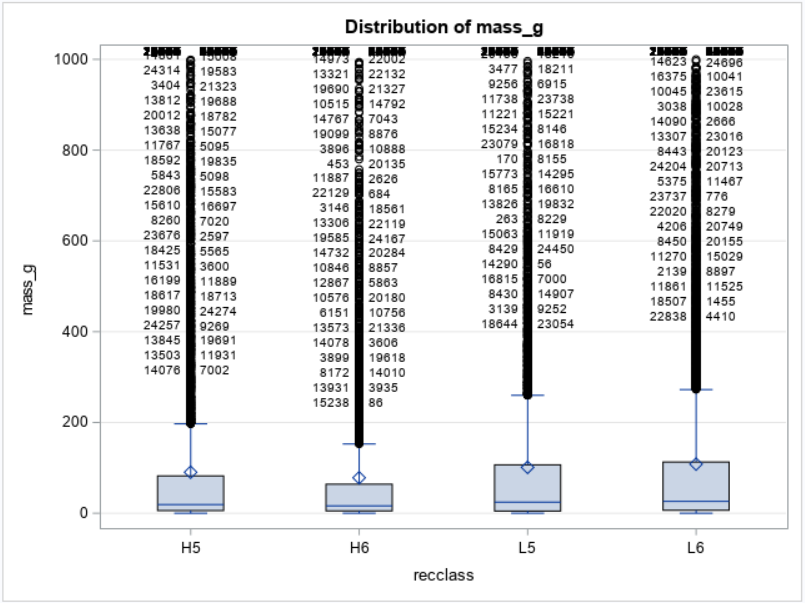
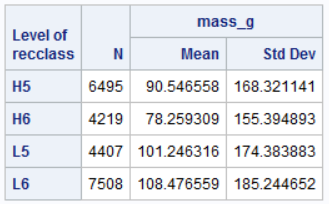


Table 5



**Interpretation of Meteorite Mass by Recclass One-way ANOVA Results**

The p-value found in the “Model” row and “Pr > F” column of table 2 was found to be < .0001. Because this value is less than .05, we reject the null hypothesis H0:

Avg mass (L6) = Avg mass (H5) = Avg mass (L5) = Avg mass (H6)

Therefore, there is sufficient evidence that the variable “recclass” in this dataset for these 4 types of asteroids the masses are not all equal. We therefore accept the alternate hypothesis HA:

The average masses of L6, H5, L5, and H6 meteorites are not all equal.

An important distinction to make in this situation is that this type of test, One-way ANOVA, does not determine which of the options of the independent variable is/are different from the others. Post-hoc testing must be implemented in order to determine which options are different and to what extent (Laerd). If the data met the assumption of homogeneity of variances, it is appropriate to use Tukey's honestly significant difference (HSD) post hoc test. If the data did not meet the homogeneity of variances assumption, it is appropriate to consider running the Games Howell post hoc test. This level of analysis will not be performed in this report.

**Part 2: Methods for Bivariate Data – Chi-Squared Test of Independence**

The Chi-squared test of independence is a statistical method which can be implemented on bivariate categorical data. The “bivariate” nature of this data to be considered is the relationship between the categorical variables “recclass” and “fall.” The dataset utilized for this section has been similarly limited to the four most common instances of the variable “recclass,” H5, H6, L5 and L6. The variable “fall” is binary in nature in that it has only two possibilities, fell and found.

If the variables “recclass” and “fall” were independent and unrelated, one might expect that as the sample grows larger that the true values of the quantities of the permutations would approach their expected values. This exercise will use the chi-squared statistic to quantify the difference of the observed and expected values of these combinations of factors. This process will be described in more detail in the following pages.

The null hypothesis for this test H0:

The variables “recclass” and “fall” are independent and unrelated.

The alternate hypothesis HA:

There is some association between the variables “recclass” and “fall.”

These hypotheses will be tested at a significance level of 0.05. There are three conditions to satisfy before using this method. [1] That the data consists of a random sample, which we will assume. [2] That the expected value for any of the data points is greater than or equal to 5, which happens to be true. The smallest expected value for this test was determined to be ~107.813. [3] The condition of independence – That the sample size consists of no more than 10% of the total population of meteorites which meet the condition to be included in this dataset. This will be assumed to be true, considering that, according to the Planetary Science Institute, more than 500 meteorites reach the surface of the Earth each year (Dhornisher). Considering the age of our planet, there are certainly more than 10 times as many meteorites as there are samples in this dataset. The following SAS code was ran in order to complete this analysis:

**proc** **freq** data = project.anova;

tables recclass\*fall / chisq;

**run**;

The first step to performing the Chi-square test of independence is to create a table which includes the two variables as rows and columns, with column totals, row totals, and a complete total. This information, along with the associated percentages, column percentages, row percentages, and total percentages, are found in Table 6 below. This information can be utilized to calculate the expected value, and subsequently the chi-squared statistic for each combination or cell. The formula for the expected value is:

Expected Value = Total number of observations \* column percentage \* row percentage

The formula for the chi-squared statistic is:

χ2= Σ [ ( Observed Value – Expected Value )2 / Expected Value ]

Hand calculated values of [ ( Observed Value – Expected Value )2 / Expected Value ] for each cell can be found in Table 7. The overall chi-squared value calculated using SAS can be found in Table 8.

Table 6

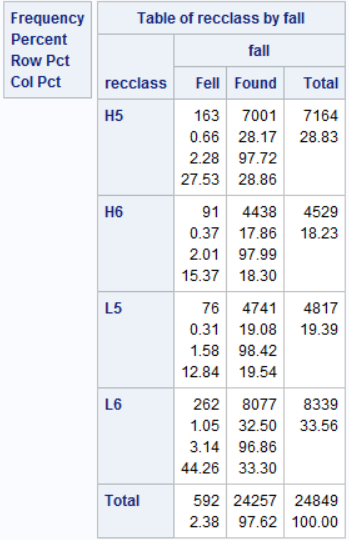
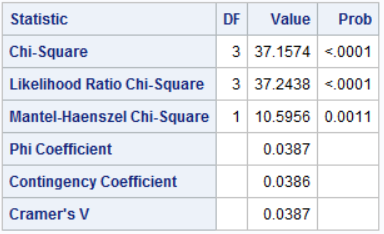


Table 7

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Fell** | **Found** |
| **H5** | Observed Value  Expected Value  Chi-Squared Statistic | 163  170.502  0.330 | 7001  6993.464  0.008 |
| **H6** | Observed Value  Expected Value  Chi-Squared Statistic | 91  107.813  2.622 | 4438  4422.159  0.057 |
| **L5** | Observed Value  Expected Value  Chi-Squared Statistic | 76  114.674  13.043 | 4741  4703.547  0.2982 |
| **L6** | Observed Value  Expected Value  Chi-Squared Statistic | 262  198.476  20.332 | 8077  8140.848  0.501 |

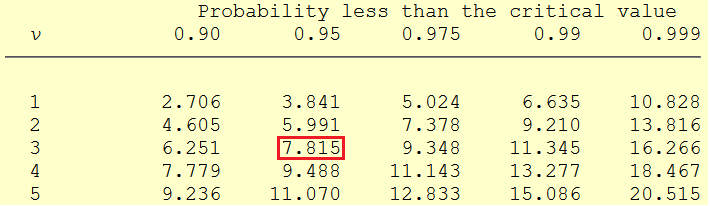
Using the values found in Table 7, the hand-calculated overall chi-squared statistic was found to be 37.19.

Table 8



SAS determined the chi-squared statistic for this dataset to be ~37.16, which is very close to the hand-calculated value of 37.19. The next step in the chi-squared test of independence is to determine the degrees of freedom. This can be determined by finding the product of the number of columns -1 and the number of rows -1, or (4 - 1) \* (2 - 1) = 3 degrees of freedom. This can also be found using SAS, which is located in the first row of Table 8.

Table 9 (1.3.6.7.4.)



At a significance level of 0.05, corresponding to a 95% confidence level, and 3 degrees of freedom, a chi squared table yields a value of 7.815. The value calculated by hand and using sas is substantially larger than this, therefore the null hypothesis will be rejected, and we will accept the alternate hypothesis, that there is some association between the variables “recclass” and “fall.” The p-value determined for this test using SAS which can be found in Table 8 is < 0.0001 and supports this conclusion.

**Conclusion**

The results of the analyses performed in this report can be summarized as follows: Utilizing a one-way ANOVA, it was determined that the average masses of meteorites for the 4 most common variants of “recclass” were not all equal, with 95% confidence. It was also determined, using a chi-squared test of independence, that the variables “recclass” and “fall” were not independent and therefore it can be said with 95% confidence that these variables are dependent on some level. These conclusions are pieces of a larger story which could be told with this data. Further analysis could be performed to explore further relationships between variables and to paint a more complete picture of meteor findings. The variables “reclat” and “reclong” were not utilized in this analysis. These variables were utilized to plot the findings on a map of the world which can be found in the appendix. It would be worthwhile and interesting to normalize the findings according to percentage of land area by latitude and longitude, as well as human population by land area in latitude and longitude. This information could be utilized to draw conclusions about whether there are some regions which are struck by meteors more frequently than others, or whether meteors are found more frequently in areas where there are more people to find them. Some effort was made to implement this type of analysis, although sadly, this depth of analysis is beyond the scope of this course.

References

Meteorite Landings. (2015, April 2). Retrieved December 10, 2019, from https://data.nasa.gov/Space-Science/Meteorite-Landings/gh4g-9sfh.

One-way ANOVA. (n.d.). Retrieved December 11, 2019, from https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php.

Dhornisher. (2019, April 16). FAQ - Meteoroids/Meteorites. Retrieved from https://www.psi.edu/epo/faq/meteor.html.

1.3.6.7.4. Critical Values of the Chi-Square Distribution. (n.d.). Retrieved December 12, 2019, from https://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm.

**APPENDIX A – COMPLETE SAS CODE**

libname project "C:\Users\conno\OneDrive\Desktop\STAT 7100 - Statistical Methods\Project";

**run**;

**data** project.data;

infile "C:\Users\conno\OneDrive\Desktop\STAT 7100 - Statistical Methods\Project\Meteorite\_Landings\_altered.csv"

delimiter = ','

MISSOVER

DSD

lrecl=**32767**

firstobs=**2**;

input name $ id nametype $ recclass $ mass\_g fall $ year $ reclat reclong GeoLocation $;

**run**;

**proc** **univariate** data = project.data;

histogram id mass\_g reclat reclong;

**run**;

**proc** **freq** data = project.data;

tables Name Nametype recclass fall;

**run**;

**data** project.data;/\*format year as numeric date\*/

set project.data;

date = input(year, mmddyy10.);

**run**;

**data** project.data;/\*format year as year\*/

set project.data;

format date year.;

**run**;

**proc** **univariate** data = project.data;

histogram date;

**run**;

/\*ordered frequency tables for all non-continuous variables\*/

**proc** **freq** data = project.data ORDER=FREQ;

tables Name / OUT = project.namefreq;

**run**;

**proc** **freq** data = project.data ORDER=FREQ;

tables Nametype / out = project.typefreq;

**run**;

**proc** **freq** data = project.data ORDER=FREQ;

tables recclass / out = project.classfreq;

**run**;

**proc** **freq** data = project.data ORDER=FREQ;

tables fall / out = project.fallfreq;

**run**;

**proc** **sort** data = project.data;

by date;

**run**;

**data** project.anova;

set project.data;

if recclass in ("L6" "H5" "L5" "H6");

**run**;

**proc** **ANOVA** data=project.anova;

title One-way ANOVA Comparing the Masses of L6, H5, L5, and H6 Meteorites;

class recclass;

model mass\_g = recclass;

means recclass;

**run**;

**proc** **freq** data = project.data;\* order=freq;

tables date;

**run**;

ods graphics on;

**proc** **freq** data=project.anova;

tables mass\_g\*recclass;\* / plots=(freqplot(scale=percent));

**run**;

ods graphics off;

**proc** **reg** data = project.data;

model mass\_g = date;

**run**;

**proc** **univariate** data = project.data;

histogram mass\_g / exponential (theta = est);

**run**;

**proc** **sort** data = project.data;

by mass\_g;

**run**;

**proc** **print** data = project.data;

var mass\_g;

**run**;

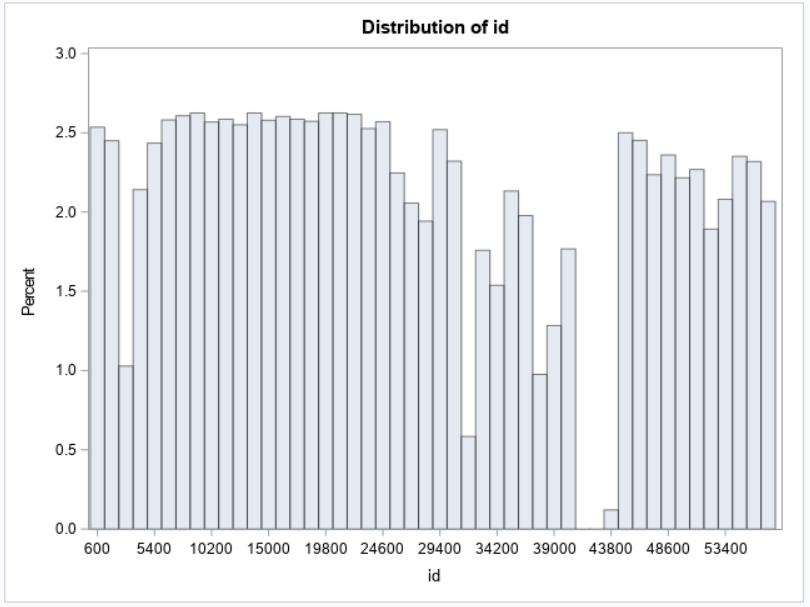
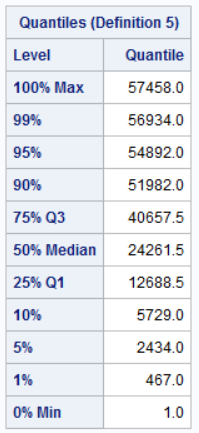
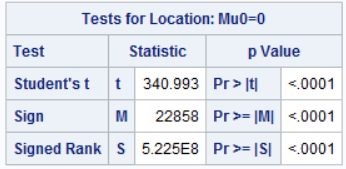
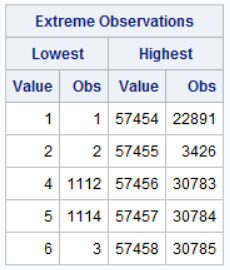
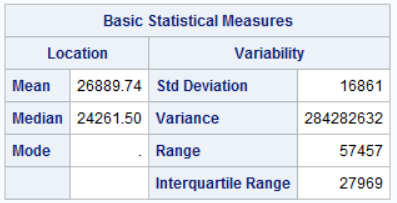
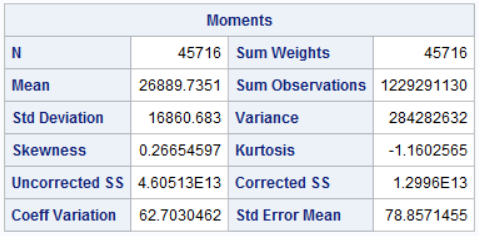
**proc** **freq** data = project.anova;

tables recclass\*fall / chisq;

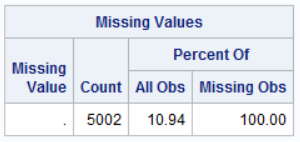
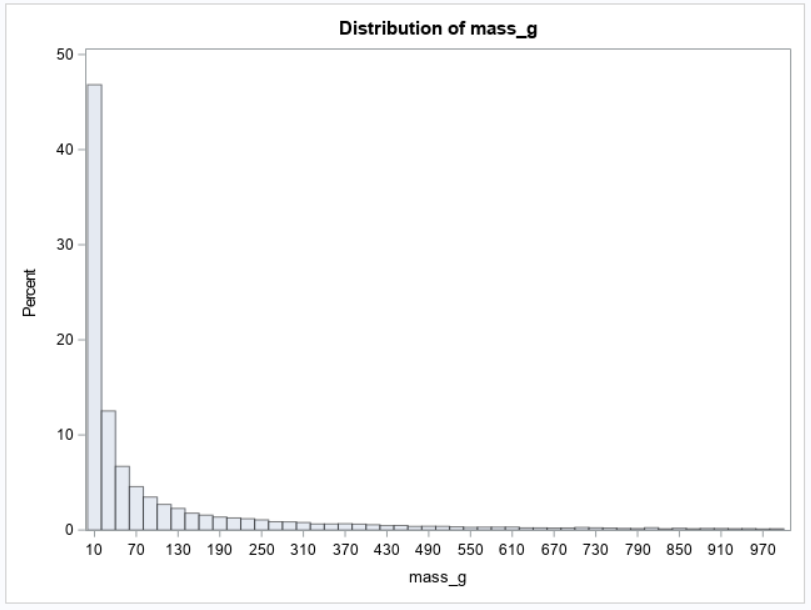
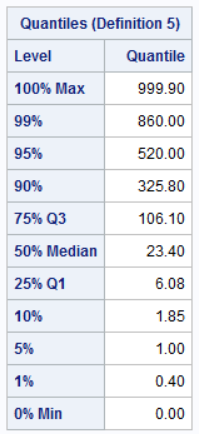
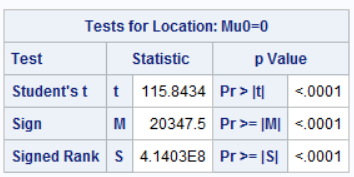
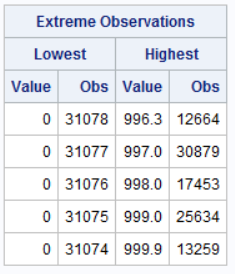
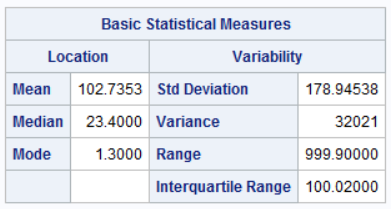
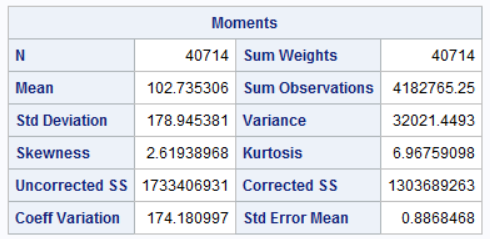
**run**;

**APPENDIX B – DESCRIPTIVE STATISTICS FOR NUMERIC VARIABLES**

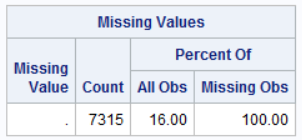
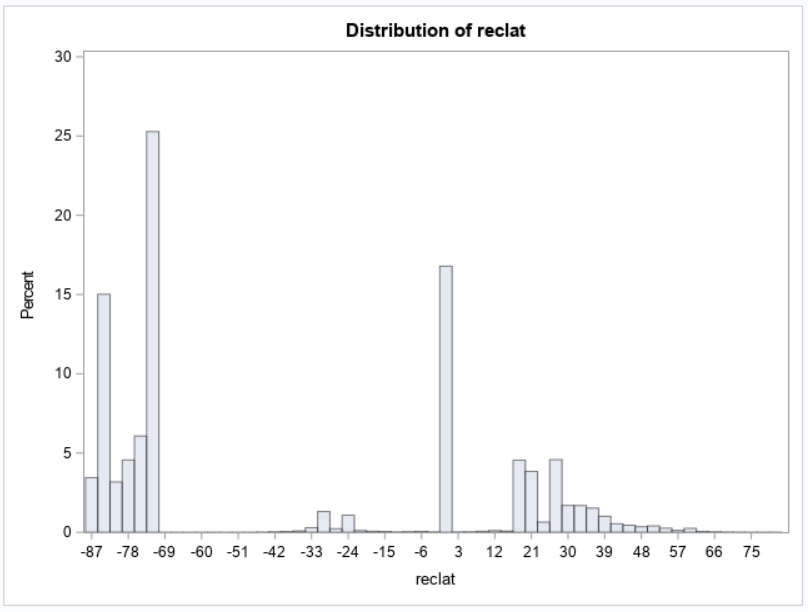
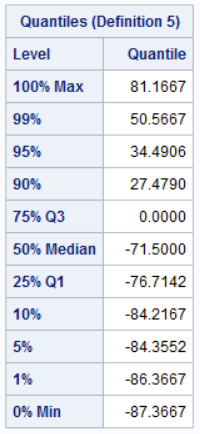
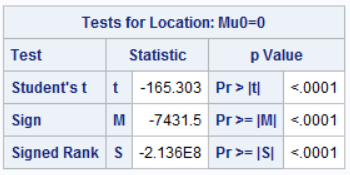
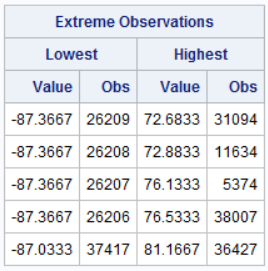
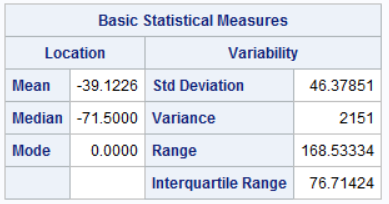
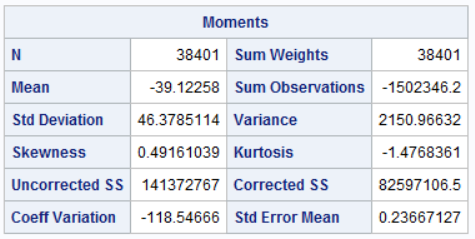
THE UNIVARIATE PROCEDURE – VARIABLE: ID



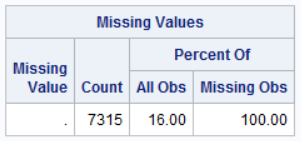
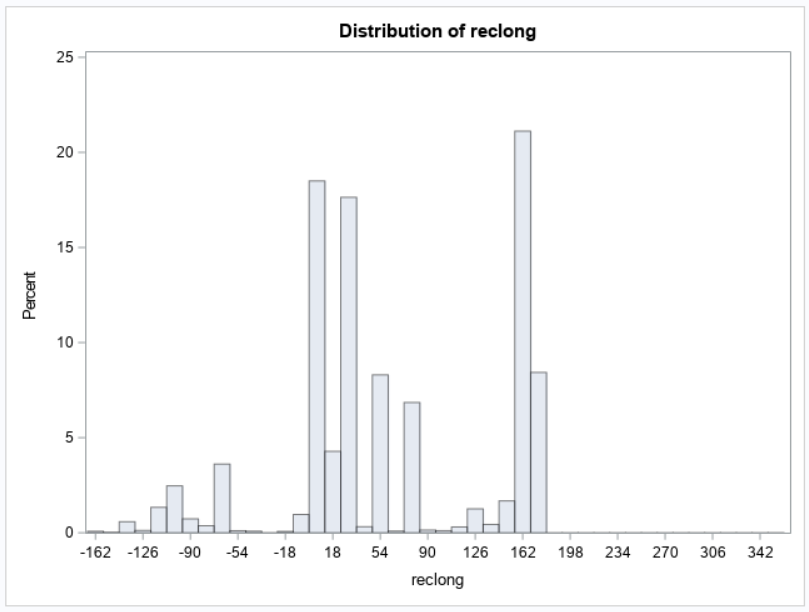
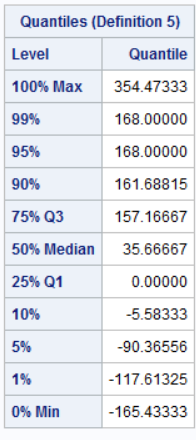
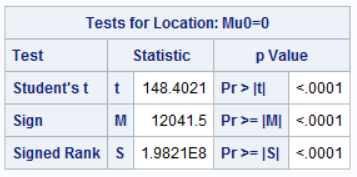
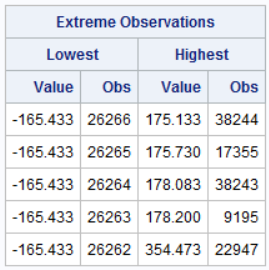
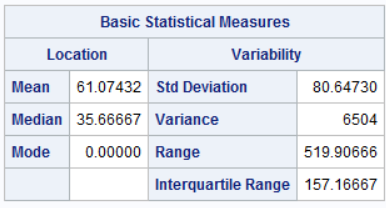
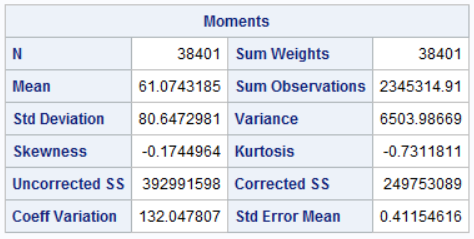
THE UNIVARIATE PROCEDURE – VARIABLE: MASS\_G



THE UNIVARIATE PROCEDURE – VARIABLE: RECLAT

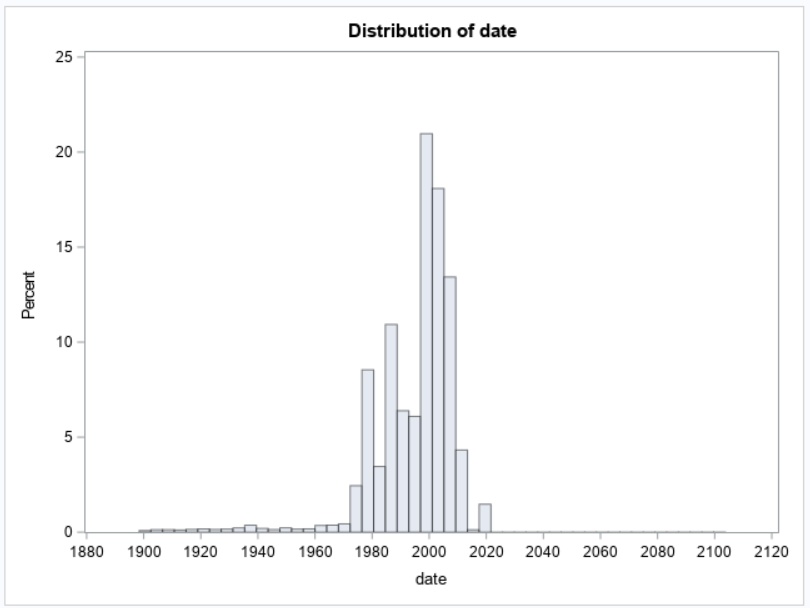
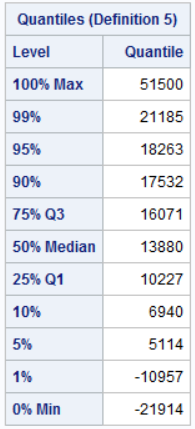
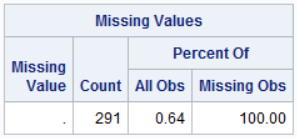
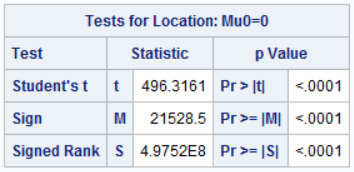
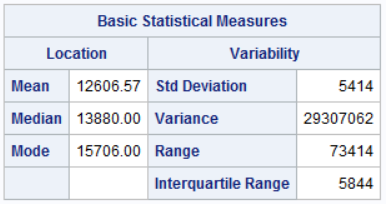
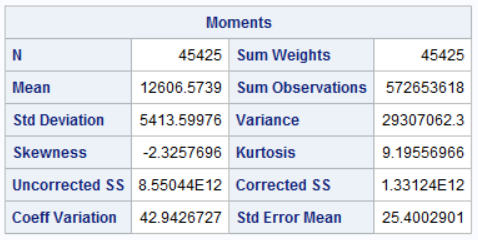


THE UNIVARIATE PROCEDURE – VARIABLE: RECLONG

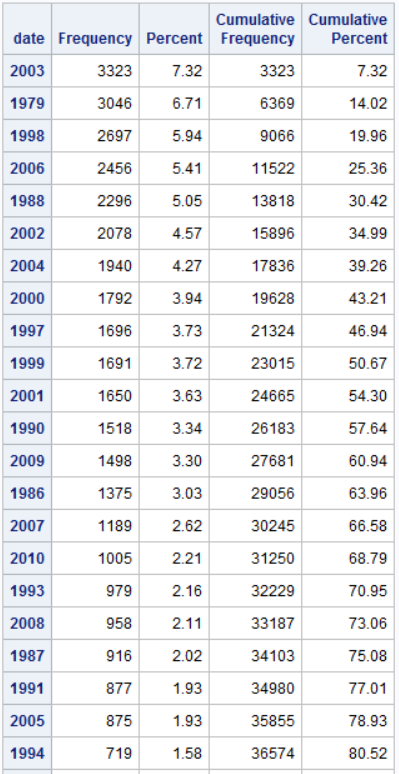


PROC UNIVARIATE PROCEDURE – VARIABLE: DATE

Notes: 291 missing values. One year value of 2101 clearly in error.



ORDERED FRQUENCY TABLE: DATE

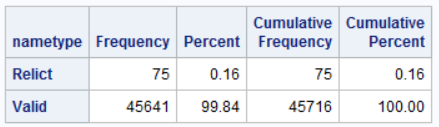


**APPENDIX C: FREQUENCY TABLES FOR NON-NUMERIC VARIABLES**

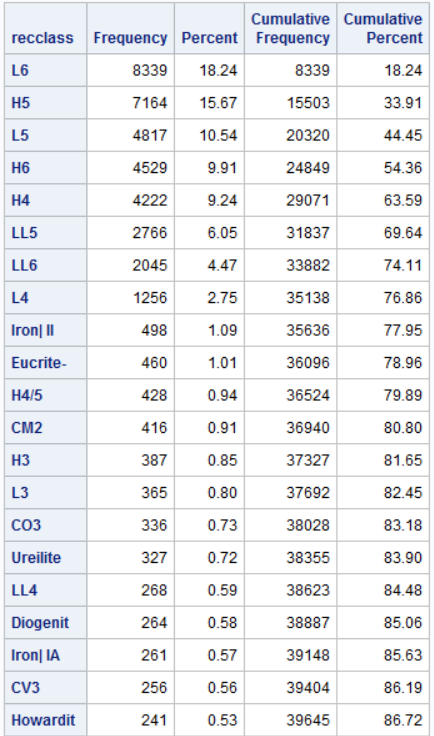
Name



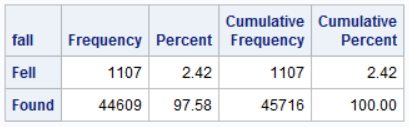
Nametype



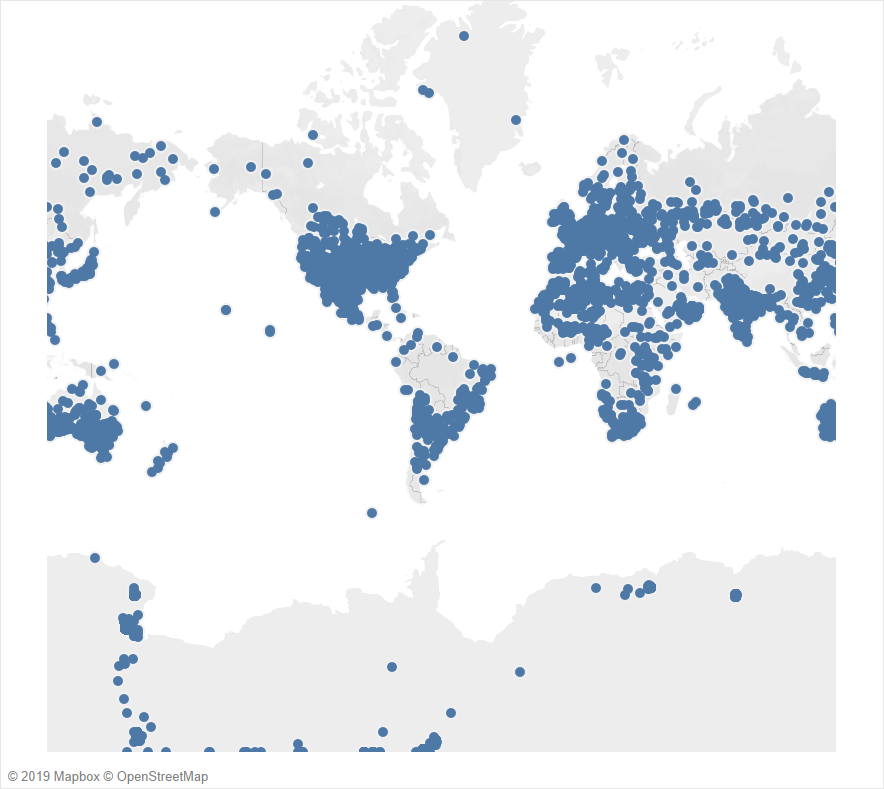
Recclass



Fall

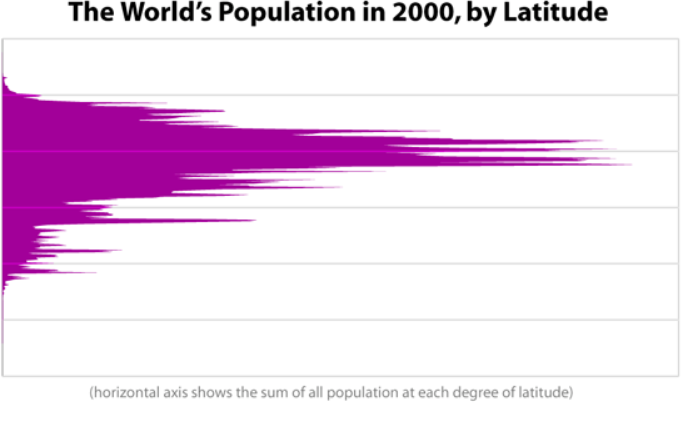


**APPENDIX D: MAP OF ALL METEORS LOCATED IN PROVIDED DATASET**

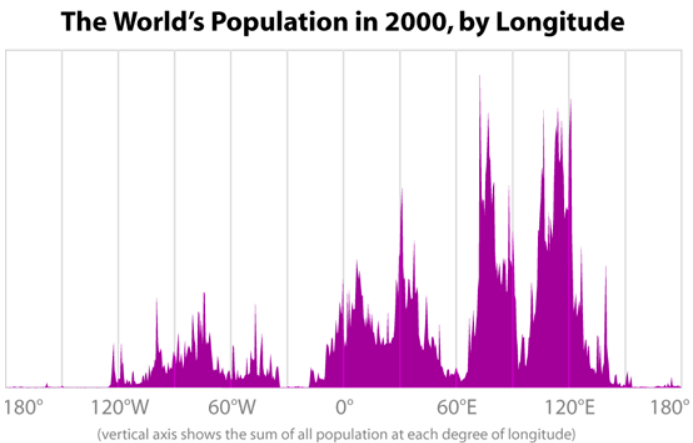


**APPENDIX E: INTERESTING MAPS BY LATITUDE AND LONGITUDE**

Future research question: are meteors found in latitudes/longitudes where there are more people?

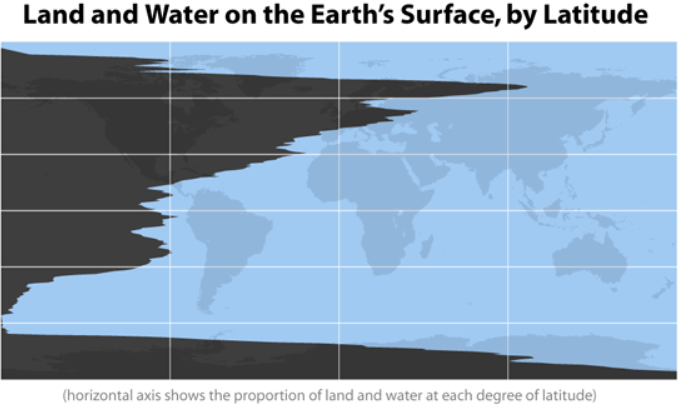


Bill Rankin, 2008, http://www.radicalcartography.net/index.html?histpop

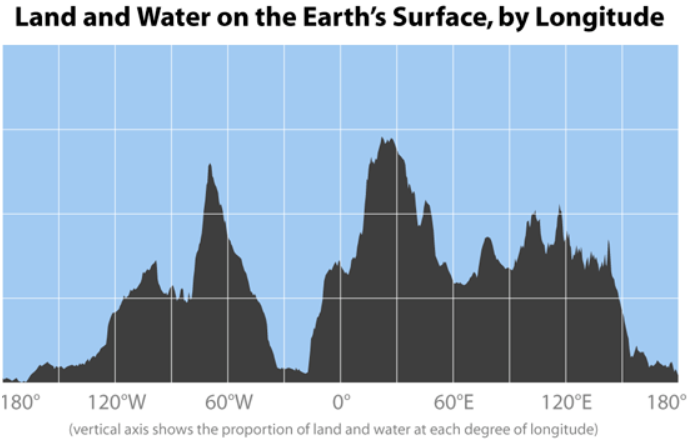


Bill Rankin, 2008, http://www.radicalcartography.net/index.html?histpop

Future research question: are meteors found in latitudes/longitudes where there is more land?



Bill Rankin, 2008, http://www.radicalcartography.net/index.html?histland



Bill Rankin, 2008, http://www.radicalcartography.net/index.html?histland