by Ezgi Polat

Ulaanbaatar Air Quality Report

Data Screening/Cleaning with SAS

1. Master Dataset and Categorization

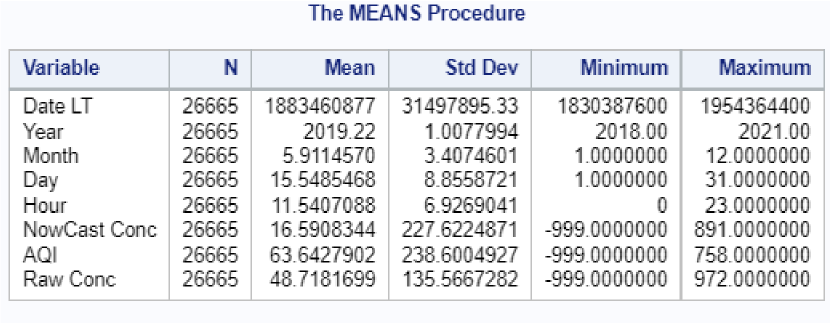
We have 4 different datasets that contain 14 columns and different numbers of rows from the years 2018-2021. Firstly, I started by assigning short names for all datasets by using %LET statement (See the code). I named the datasets as dataset_2018, dataset_2019, dataset_2020, and dataset_2021. Then I used Proc Import with dbms, out, replace and guessingrow statements to import the data to SAS.

We can see that our 2018 dataset contains 7936 rows, the 2019 dataset contains 8329 rows, the 2020 dataset contains 7065 rows, and the 2021 dataset contains 3335 rows. I created a master dataset to merge all 4 datasets and named it '**masterdata**' which contains a total of **26,665 rows** and **14 columns**.

We have a combination of numeric and categorical data in our dataset. Here is the categorization of our variables;

Site	Categorical
Parameter	Categorical
Date	Numerical
Year	Numerical
Month	Numerical
Day	Numerical
Hour	Numerical
NowCast Conc	Numerical
AQI	Numerical
AQI_Category	Categorical
RowConc	Numerical
Conc Unit	Categorical
Duration	Categorical
QC Name	Categorical

We have **8 numerical** and **6 categorical** variables in our dataset. To confirm if SAS processed the right variables as numerical, I used Proc Means statement and checked which variables were processed as numerical. In the result panel (Image 1), we can see that 8 of the variables were processed as numerical which are Date LT, Year, Month, Day, Hour, NowCast Conc, AQI, and Raw Conc. Therefore, I confirmed that SAS categorized our variables correctly.



The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Date LT	26665	1883460877	31497895.33	1830387600	1954364400
Year	26665	2019.22	1.0077994	2018.00	2021.00
Month	26665	5.9114570	3.4074601	1.0000000	12.0000000
Day	26665	15.5485468	8.8558721	1.0000000	31.0000000
Hour	26665	11.5407088	6.9269041	0	23.0000000
NowCast Conc	26665	16.5908344	227.6224871	-999.0000000	891.0000000
AQI	26665	63.6427902	238.6004927	-999.0000000	758.0000000
Raw Conc	26665	48.7181699	135.5667282	-999.0000000	972.0000000

Image 1. Proc Means Results Showing 8 Numerical Categories

2. Accuracy

Proc Means statement (see appx 1) also helped me to see some out-of-range scores (Image 1). **NowCast Conc**, **AQI** and **Raw Conc** variables has a minimum value of -999; and maximum values of 891, 758 and 972. According to the Fact Sheet File, AQI levels are supposed to be **between 0 and 500**. This means we have inaccurate data entries for these variables. Any data lower than 0 and higher than 500 should be considered as missing data.

Other numerical variables, such as Date, Year, Month, Day and Hour, have accurate values. There are no accuracy issues for their min and max values.

I also checked the histogram of these 3 variables (see appx 2) to have a better understanding of the out-of-range scores.

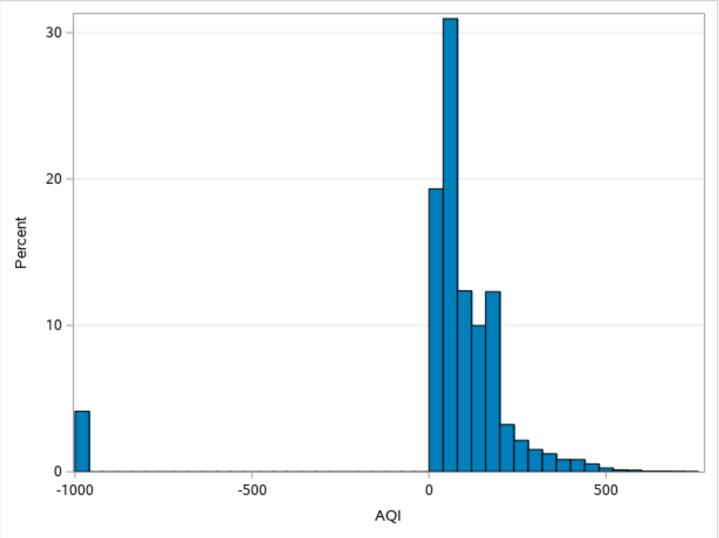


Image 2. The Histogram of AQI Variable

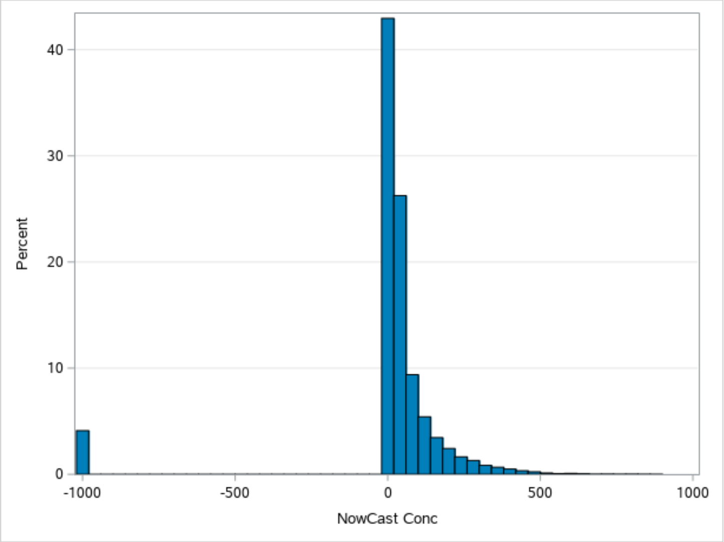


Image 3. The Histogram of NowCast Conc Variable

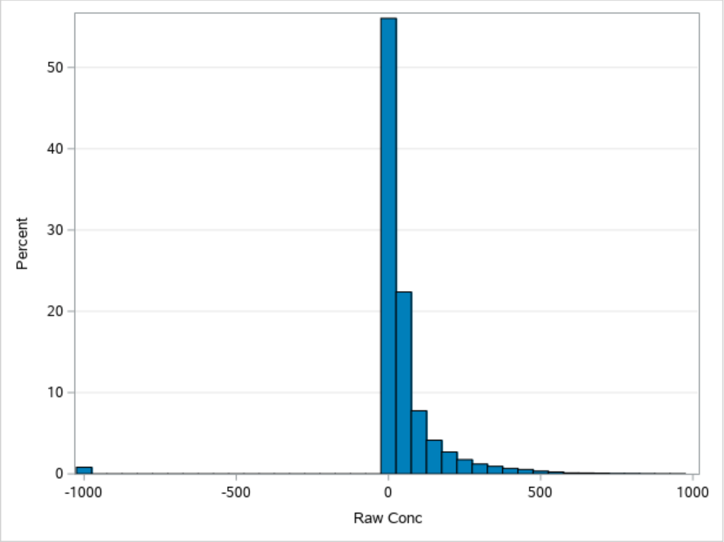


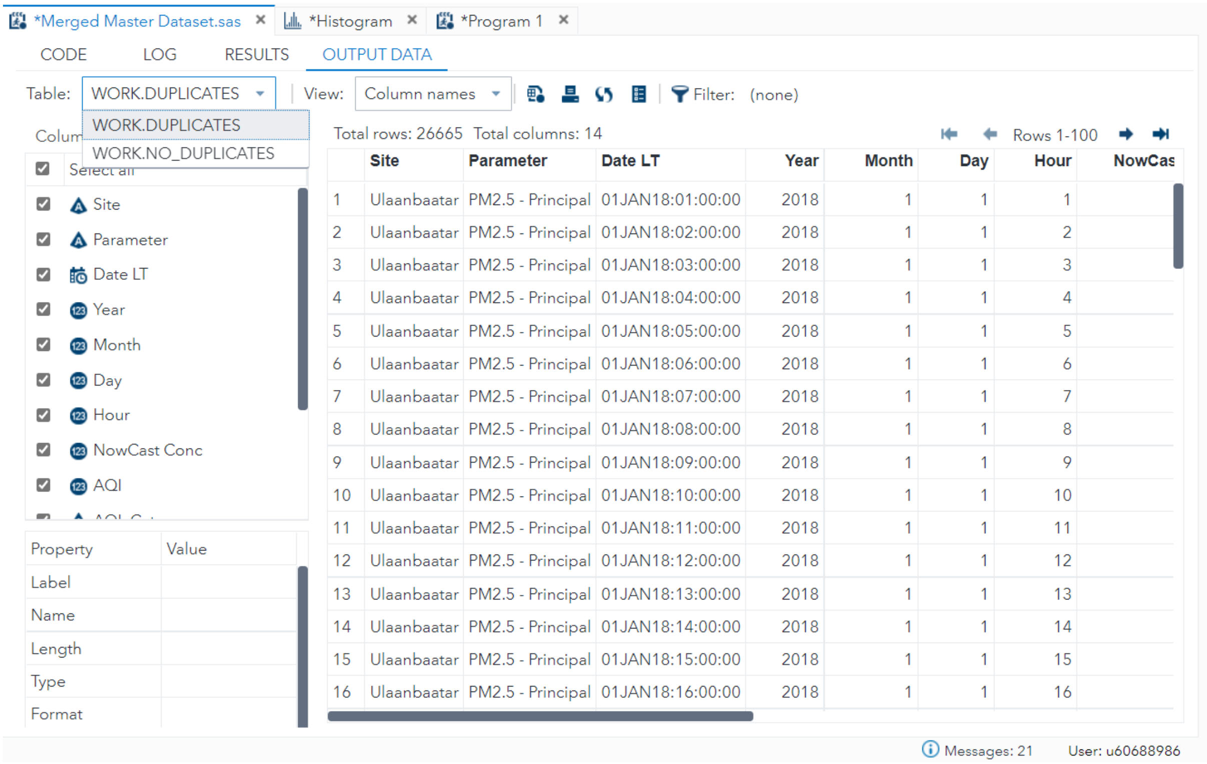
Image 4. The Histogram of Raw Conc Variable

We can see the out-of-range scores in histograms as well. The percentage of these scores are **less than 5%** of the masterdata in all variables (NowCast Conc; Raw Conc and AQI).

To resolve the issue, my plan is to label all out-of-range values as (‘’) in this step and then remove all (‘’) values in the next step. But before doing that; I wanted to check my dataset for **duplicates**.

I tried 4 different codes to identify/remove the duplicates (see appx 3). But I did not identify any duplicates with any of the codes. The row numbers always remained the same.

One of the ways I used to identify duplicates was to create 2 separate output tables showing unique and duplicate observations (see appx 3.D). Work_Duplicates dataset in the Image 5 shows the **unique** observations. It has 26665 rows which is equal to the masterdata dataset.



*Merged Master Dataset.sas *Histogram *Program 1

CODE LOG RESULTS OUTPUT DATA

Table: WORK.DUPLICATES View: Column names Filter: (none)

WORK.DUPLICATES
WORK.NO_DUPLICATES

Columns

☑ Select all
☑ Site
☑ Parameter
☑ Date LT
☑ Year
☑ Month
☑ Day
☑ Hour
☑ NowCast Conc
☑ AQI

Property	Value
Label	
Name	
Length	
Type	
Format	

Total rows: 26665 Total columns: 14

Rows 1-100

	Site	Parameter	Date LT	Year	Month	Day	Hour	NowCas
1	Ulaanbaatar	PM2.5 - Principal	01JAN18:01:00:00	2018	1	1	1	
2	Ulaanbaatar	PM2.5 - Principal	01JAN18:02:00:00	2018	1	1	2	
3	Ulaanbaatar	PM2.5 - Principal	01JAN18:03:00:00	2018	1	1	3	
4	Ulaanbaatar	PM2.5 - Principal	01JAN18:04:00:00	2018	1	1	4	
5	Ulaanbaatar	PM2.5 - Principal	01JAN18:05:00:00	2018	1	1	5	
6	Ulaanbaatar	PM2.5 - Principal	01JAN18:06:00:00	2018	1	1	6	
7	Ulaanbaatar	PM2.5 - Principal	01JAN18:07:00:00	2018	1	1	7	
8	Ulaanbaatar	PM2.5 - Principal	01JAN18:08:00:00	2018	1	1	8	
9	Ulaanbaatar	PM2.5 - Principal	01JAN18:09:00:00	2018	1	1	9	
10	Ulaanbaatar	PM2.5 - Principal	01JAN18:10:00:00	2018	1	1	10	
11	Ulaanbaatar	PM2.5 - Principal	01JAN18:11:00:00	2018	1	1	11	
12	Ulaanbaatar	PM2.5 - Principal	01JAN18:12:00:00	2018	1	1	12	
13	Ulaanbaatar	PM2.5 - Principal	01JAN18:13:00:00	2018	1	1	13	
14	Ulaanbaatar	PM2.5 - Principal	01JAN18:14:00:00	2018	1	1	14	
15	Ulaanbaatar	PM2.5 - Principal	01JAN18:15:00:00	2018	1	1	15	
16	Ulaanbaatar	PM2.5 - Principal	01JAN18:16:00:00	2018	1	1	16	

Messages: 21 User: u60688986

Image 5. Work_Duplicates Dataset

Work_No_Duplicates dataset in Image 6 shows the **duplicate** observations, and it has 0 observation. Therefore, we can say that there are no accuracy issues in masterdata about duplicates.

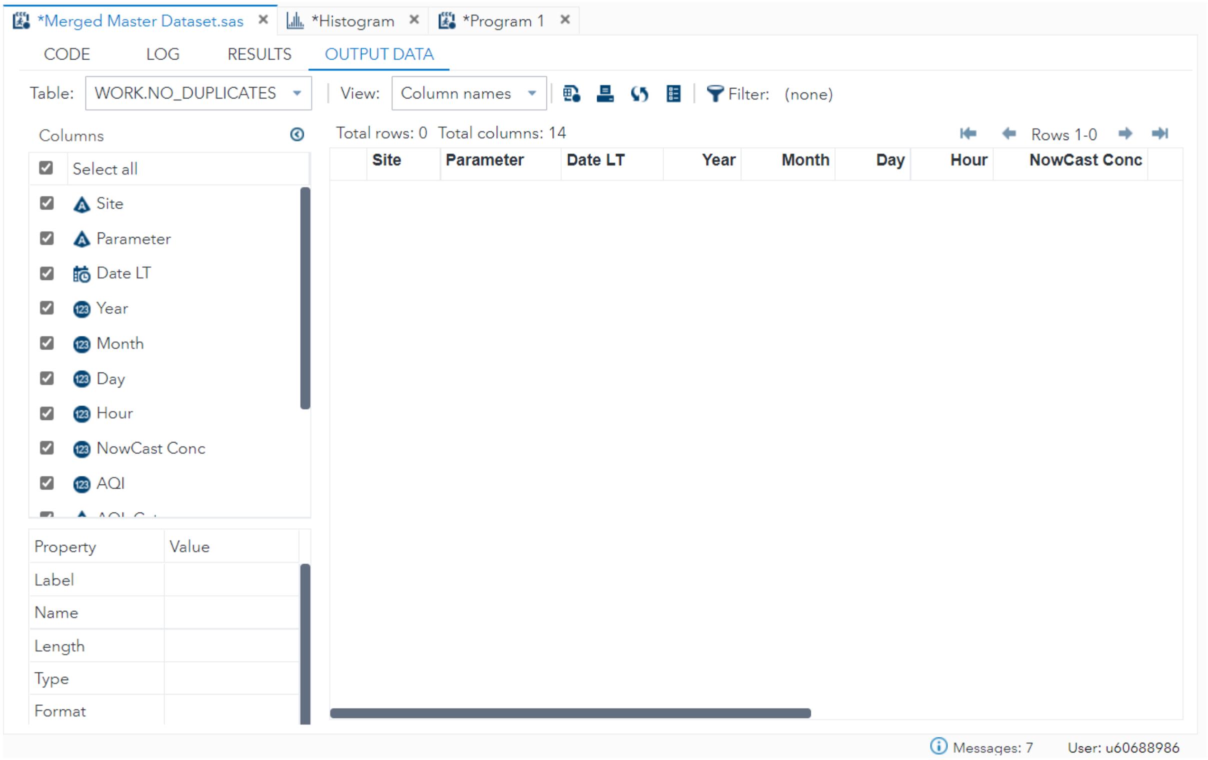


Table: WORK.NO_DUPLICATES View: Column names Filter: (none)

Columns

☑ Select all
☑ Site
☑ Parameter
☑ Date LT
☑ Year
☑ Month
☑ Day
☑ Hour
☑ NowCast Conc
☑ AQI

Total rows: 0 Total columns: 14 Rows 1-0

Site	Parameter	Date LT	Year	Month	Day	Hour	NowCast Conc

Property	Value
Label	
Name	
Length	
Type	
Format	

Messages: 7 User: u60688986

Image 6. Work_No_Duplicates Dataset

After this step, I went back to the out-of-range scores in NowCast Conc & AQI & Raw Conc variables to **label them as missing data/(‘’)**.

First, I wanted to see the numbers of all out-of-range scores in each variable. I used Proc Print Statement to see the numbers of out-of-range values in NowCast Conc variable (see appx 4) and there were 1221 observations (in LOG) that shows the values that are NOT between 0 and 500. I did the same for AQI (see appx 5) and there were 1221 observations again. Lastly, I tried it for Raw Conc (see appx 6) and got 1174 out-of-range observations.

Next step was to label all out-of-range scores as (‘’)/missing data. I started doing that by setting a new dataset where NowCast Conc values that are NOT between 0 and 500 is labeled as (‘’) (see appx 7) and named this new dataset as masterdata1.

As second step, I did the exact same process for AQI variable using my new masterdata1 dataset. In masterdata1, out-of-range scores of NowCast Conc was already labeled as (‘’) and now it was time for AQI variable. Hence, I created masterdata2 by labeling out-of-range scores of AQI as (‘’) in masterdata1 (see appx 8). At the end, I had masterdata2 where all out-of-range scores of NowCast Conc & AQI variables are labeled as (‘’).

Lastly, I applied the same process for Raw Conc variable using masterdata2 and as a result I created the dataset named masterdata3 where all out-of-range scores of all NowCast Conc & AQI & Raw Conc variables are labeled as missing data (see appx 9).

3. Missing Data

Now that I have a new dataset, masterdata3, where all the out-of-range scores are labeled as (‘’), I removed all rows with the **numerical missing values** from the masterdata3 with Then Delete statement (see appx 10). At the end, **I removed 1804 rows** and as it can be seen in Image 7, now my dataset has **24,861** rows and all out-of-range scores are cleaned. I named this dataset version as masterdata4.

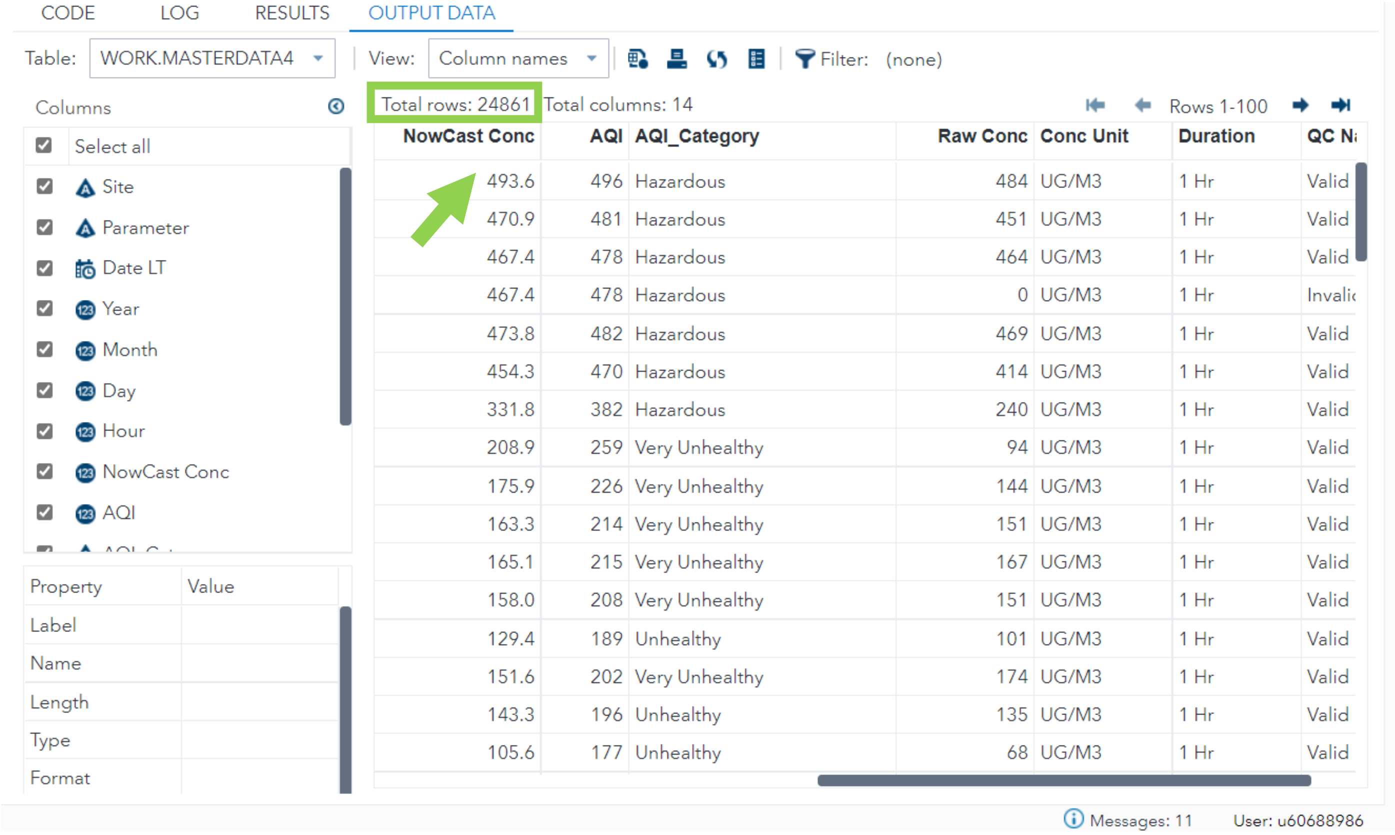


Table: WORK.MASTERDATA4 View: Column names Filter: (none)

Columns

☑ Select all
☑ Site
☑ Parameter
☑ Date LT
☑ Year
☑ Month
☑ Day
☑ Hour
☑ NowCast Conc
☑ AQI

Property	Value
Label	
Name	
Length	
Type	
Format	

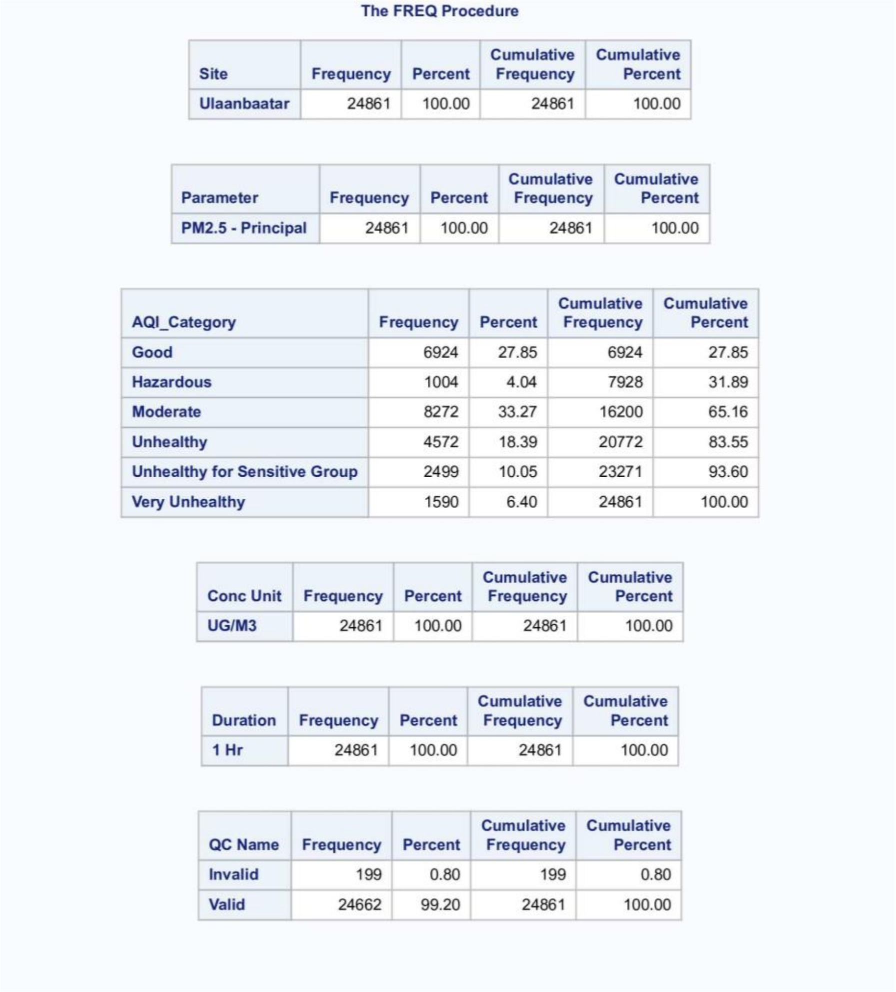
Total rows: 24861 Total columns: 14 Rows 1-100

NowCast Conc	AQI	AQI_Category	Raw Conc	Conc Unit	Duration	QC N
493.6	496	Hazardous	484	UG/M3	1 Hr	Valid
470.9	481	Hazardous	451	UG/M3	1 Hr	Valid
467.4	478	Hazardous	464	UG/M3	1 Hr	Valid
467.4	478	Hazardous	0	UG/M3	1 Hr	Invali
473.8	482	Hazardous	469	UG/M3	1 Hr	Valid
454.3	470	Hazardous	414	UG/M3	1 Hr	Valid
331.8	382	Hazardous	240	UG/M3	1 Hr	Valid
208.9	259	Very Unhealthy	94	UG/M3	1 Hr	Valid
175.9	226	Very Unhealthy	144	UG/M3	1 Hr	Valid
163.3	214	Very Unhealthy	151	UG/M3	1 Hr	Valid
165.1	215	Very Unhealthy	167	UG/M3	1 Hr	Valid
158.0	208	Very Unhealthy	151	UG/M3	1 Hr	Valid
129.4	189	Unhealthy	101	UG/M3	1 Hr	Valid
151.6	202	Very Unhealthy	174	UG/M3	1 Hr	Valid
143.3	196	Unhealthy	135	UG/M3	1 Hr	Valid
105.6	177	Unhealthy	68	UG/M3	1 Hr	Valid

Messages: 11 User: u60688986

Image 7. Cleaned Data

Then Delete statement removed all missing numerical values but I wanted to check **categorical missing values** as well by using Proc Freq statement (see appx 11).



The FREQ Procedure

Site	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Ulaanbaatar	24861	100.00	24861	100.00

Parameter	Frequency	Percent	Cumulative Frequency	Cumulative Percent
PM2.5 - Principal	24861	100.00	24861	100.00

AQI_Category	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Good	6924	27.85	6924	27.85
Hazardous	1004	4.04	7928	31.89
Moderate	8272	33.27	16200	65.16
Unhealthy	4572	18.39	20772	83.55
Unhealthy for Sensitive Group	2499	10.05	23271	93.60
Very Unhealthy	1590	6.40	24861	100.00

Conc Unit	Frequency	Percent	Cumulative Frequency	Cumulative Percent
UG/M3	24861	100.00	24861	100.00

Duration	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1 Hr	24861	100.00	24861	100.00

QC Name	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Invalid	199	0.80	199	0.80
Valid	24662	99.20	24861	100.00

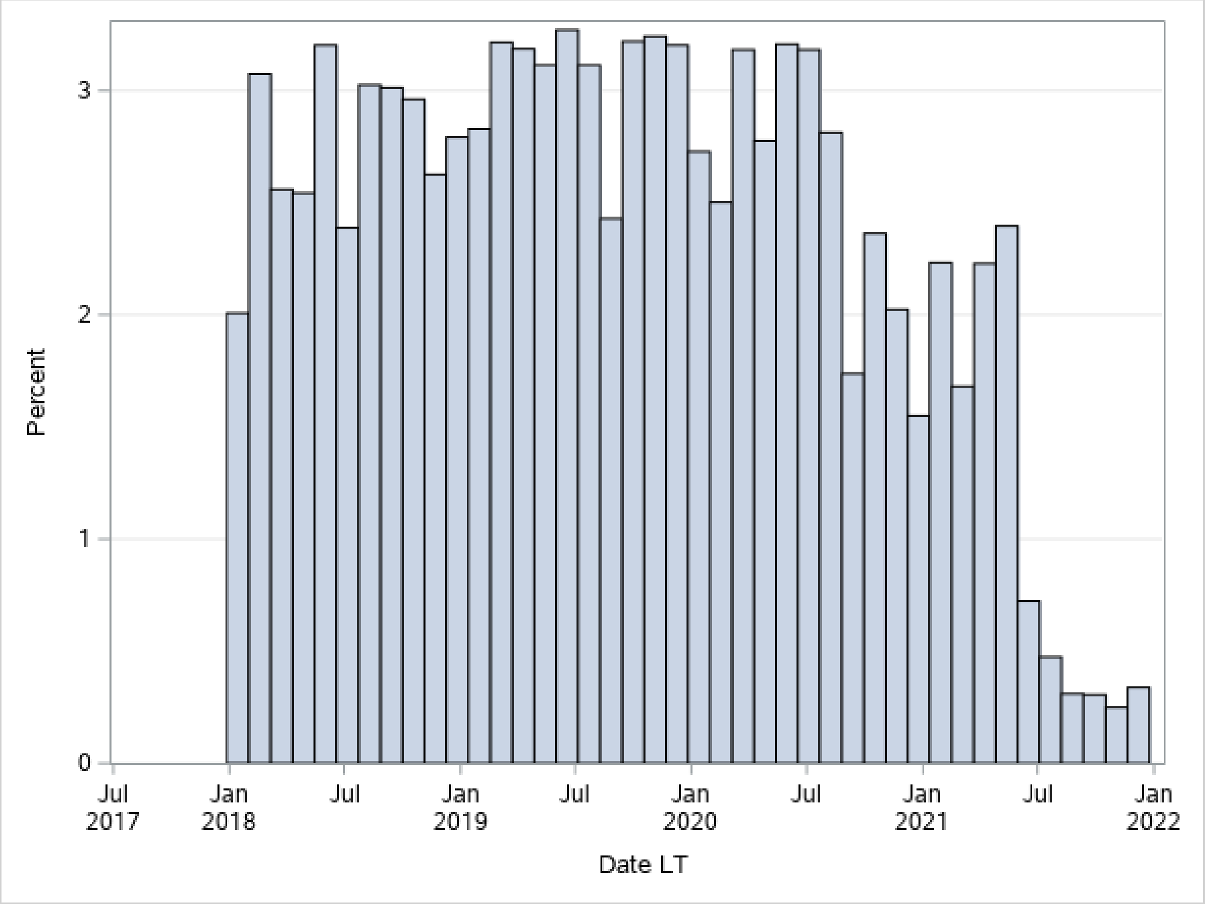
Image 8. Proc Freq Results

As it can be seen in Image 8, all categorical variables look good and there is no missing data or errors like N/A.

Lastly, I used Univariate statement (see appx 12) to double-check all values including min and max values of all variables. Min and Max values were in range of 0-500.

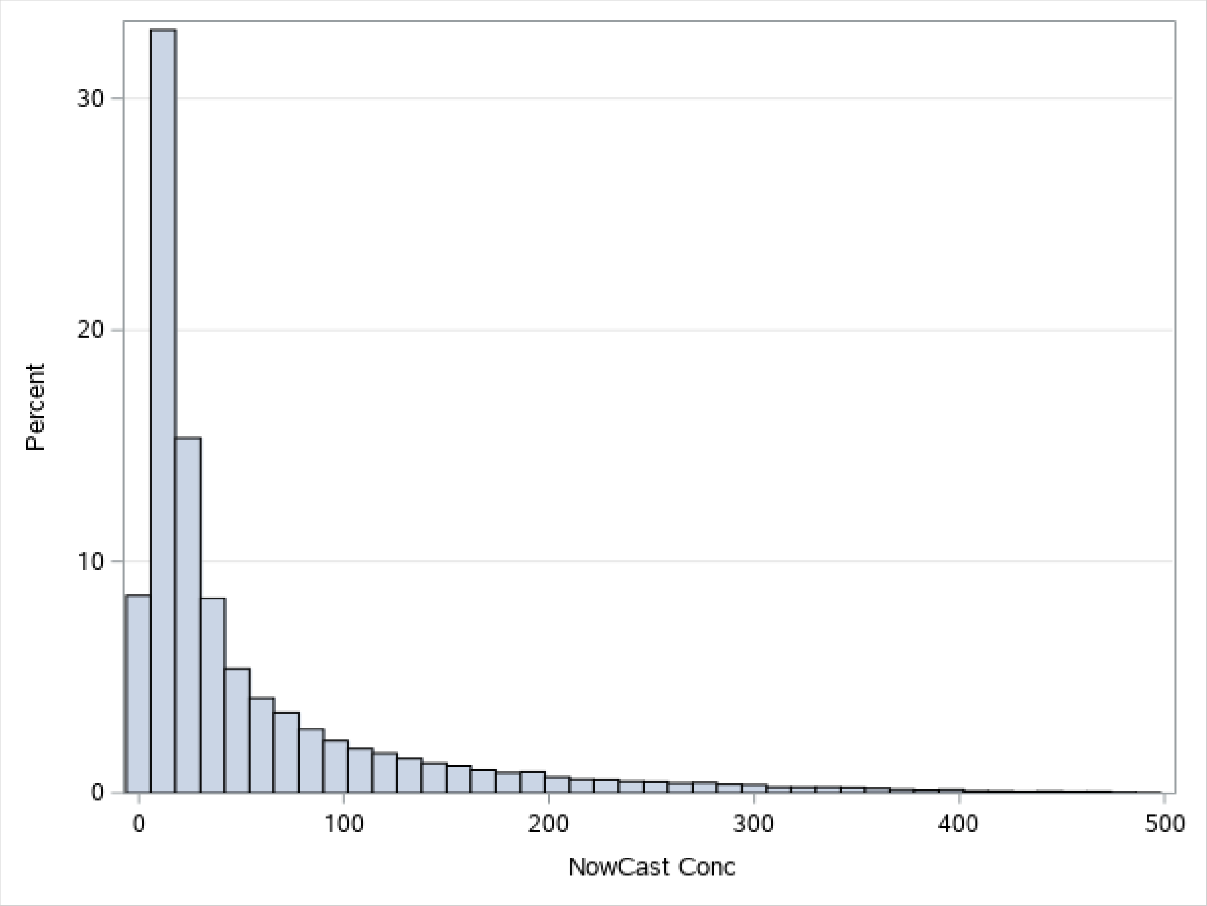
5. Univariate Normality

The Histogram of Dates;



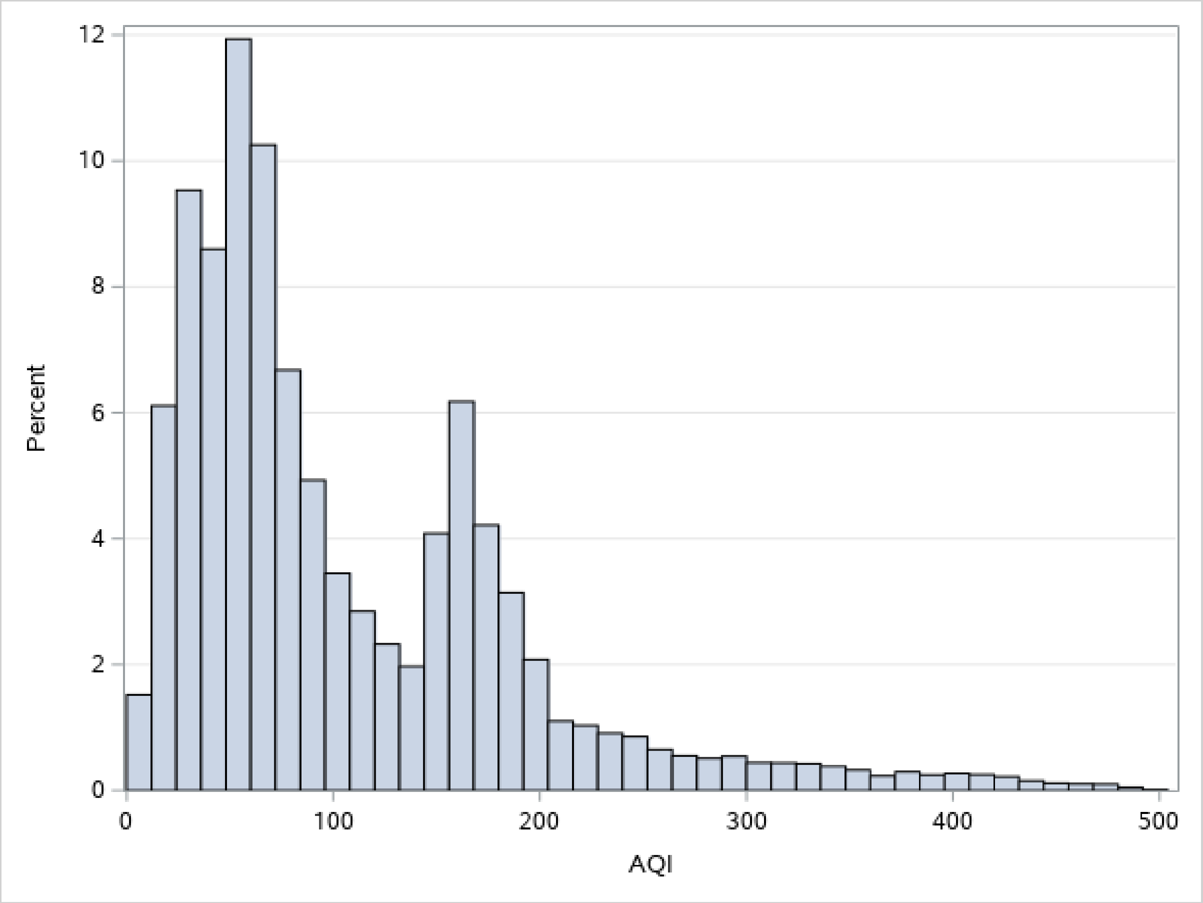
The histogram type is **UNIFORM.**

The Histogram of NowCast Conc;



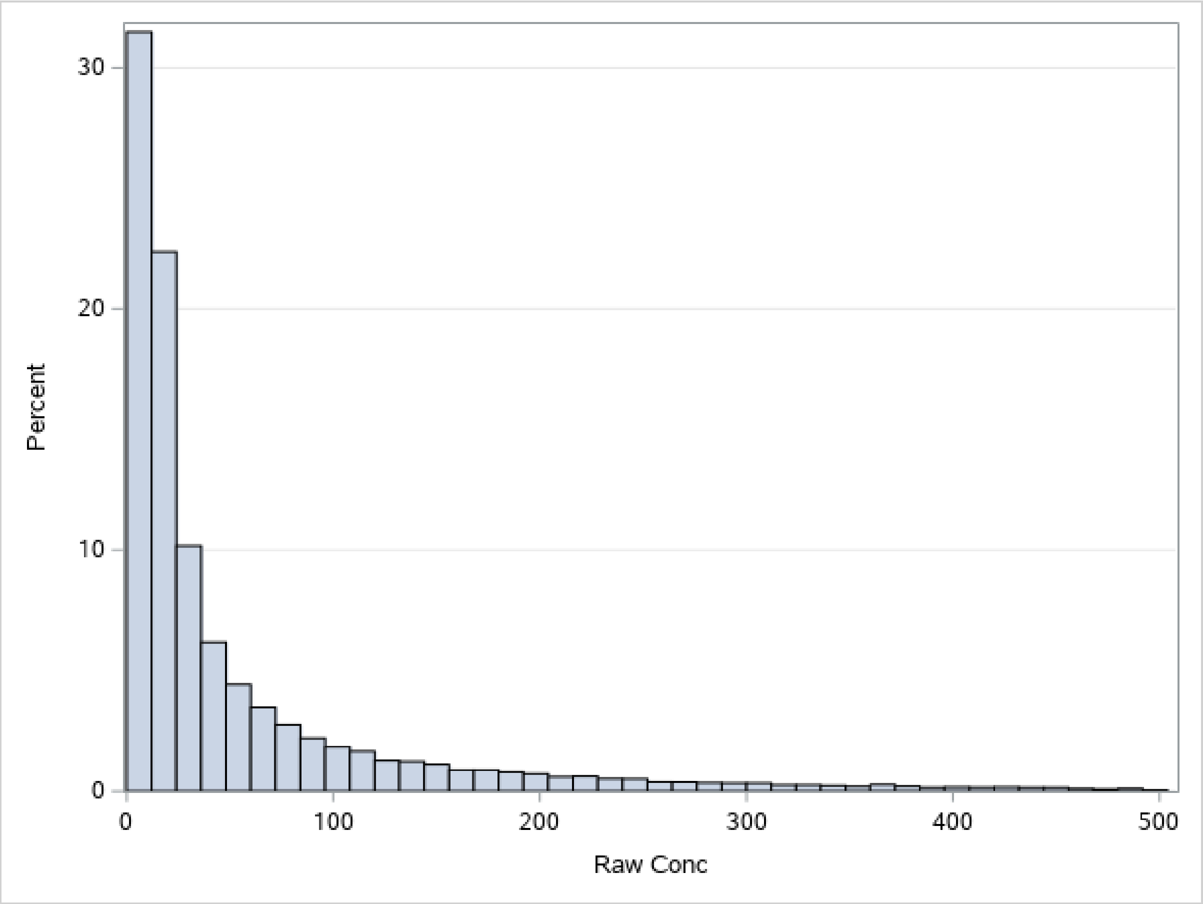
The histogram type is **SKEW RIGHT.**

The Histogram of AQI;



The histogram type is **BIMODAL**

The Histogram of Raw Conc;



The histogram type is **SKEW RIGHT.**

APPENDIX

1

```
proc means data=work.masterdata;

run;
```

2

NowCast Conc Histogram;

```
ods graphics / reset width=6.4in height=4.8in imagemap;


proc sgplot data=WORK.MASTERDATA;

	histogram 'NowCast Conc'n / fillattrs=(color=CX0e75b9);

	yaxis grid;

run;


ods graphics / reset;
```

Raw Conc Histogram;

```
ods graphics / reset width=6.4in height=4.8in imagemap;


proc sgplot data=WORK.MASTERDATA;

	histogram 'Raw Conc'n / fillattrs=(color=CX0e75b9);

	yaxis grid;

run;
```

```
ods graphics / reset;
```

AQI Histogram;

```
ods graphics / reset width=6.4in height=4.8in imagemap;


proc sgplot data=WORK.MASTERDATA;
        histogram AQI / fillattrs=(color=CX0e75b9);
        yaxis grid;
run;


ods graphics / reset;
```

3

```
A.      proc sort data=work.masterdata nodupkey;
        by _all_;
        run;



B.      proc sort data=work.masterdata nodupkey;
        by 'Date LT'n;
        run;


C.      proc sort data=work.masterdata;
        by _all_;
        run;
```

```
D.      data work.no_duplicates work.duplicates;

        set work.masterdata;

        by 'Date LT'n;


        if first.product then output work.no_duplicates;

        else output work.duplicates;

        run;
```

4

```
proc print data=work.masterdata;

WHERE 'NowCast Conc'n < 0 or 'NowCast Conc'n > 500;

run;
```

5

```
proc print data=work.masterdata;

WHERE AQI < 0 or AQI > 500;

run;
```

6

```
proc print data=work.masterdata;

WHERE 'Raw Conc'n < 0 or 'Raw Conc'n > 500;

run;
```

7

```
data masterdata1;
```

```
SET work.masterdata;

If 'NowCast Conc'n < 0 or 'NowCast Conc'n > 500  THEN 'NowCast Conc'n = '';

run;
```

8

```
data masterdata2;

SET work.masterdata1;

If AQI < 0 or AQI > 500  THEN AQI = '';

run;
```

9

```
data masterdata3;

SET work.masterdata2;

If 'Raw Conc'n < 0 or 'Raw Conc'n > 500  THEN 'Raw Conc'n = '';

run;
```

10

```
data masterdata4;

set masterdata3;

 if cmiss(of _all_) then delete;

run;
```

11

```
proc freq data=masterdata4;

tables Site Parameter AQI_Category 'Conc Unit'n Duration 'QC Name'n;

run;
```

12

```
proc freq data=masterdata4;

tables Site Parameter AQI_Category 'Conc Unit'n Duration 'QC Name'n;

run;
```