I ran it on SASHELP.ZIPCODES added a few variables to highlight some of the items above

Here are some of the interesting discoveries

1. NO need for PONAME and CITY - They are exact duplicates? (note same number unique vals is also a tipoff)

2. ZIP is the primary key

3. Using optimum lengths reduces the size of the dataset by 72 bytes (fat variables should be codes)

4. Longitude is very strongly related to numeric zipcode spearman cor of 0.96

5. Greenwich meantime is obviously related to zipcode, since it is a measure of longitude?

6. Outlier analysis through univariate

7 One to One GMTOFFSET to TIMEZONE ( may not need both of these variables)

8 One to One STATE to STATECODE (need both?)

9 One to One STATE to STATENAME (need both)

10 The strongest association between unordered catagorical variabkes is STATE \* STATECODE

11 The weakest is Rand\_A \* Rand\_B (0 to 1 scale for Cramers V like Pearson Correlation)

Character Variables with All Equal Values(only need on rog these variables)

Variable Type Len Compare Len Label Compare Label

CITY CHAR 35 PONAME 35 Name of city/org USPS Post Office Name: same as City

Dataset summary for WORK.CHRRELFIN

Observations: 41,466

Variables: 21

----------------------------------------

# Variable Unique Values Label Type Length Format

--- -------- ------------- ----- ---- ------ ------

19 ZIP 41,466 The 5-digit ZIP Code num 8 Z5. PRiMARY KEY

1 ALIAS\_CITY 6,827 USPS - alternate names of city separatedchar 300

2 ALIAS\_CITYN 4,236 Local - alternate names of city separatechar 300

3 AREACODE 247 Single Area Code for ZIP Code. num 8 close carinality

4 AREACODES 255 Multiple Area Codes for ZIP Code. char 12

5 CITY 18,705 Name of city/org char 35 Same cardinality

12 PONAME 18,705 USPS Post Office Name: same as City char 35

6 CNT 116 num 8

7 COUNTY 326 FIPS county code. num 8

8 COUNTYNM 1,931 Name of county/parish. char 25

9 DST 2 ZIP Code obeys Daylight Savings: Y-Yes Nchar 1

10 GMTOFFSET 11 Diff (hrs) between GMT and time zone fornum 8

11 MSA 338 Metro Statistical Area code by common ponum 8

13 STATE 58 Two-digit number (FIPS code) for state/tnum 8

14 STATECODE 58 Two-letter abbrev. for state name. char 2

15 STATENAME 58 Full name of state/territory char 25

16 TIMEZONE 11 Time Zone for ZIP Code. char 9

17 X 41,365 Longitude (degrees) of the center (centrnum 8

18 Y 41,317 Latitude (degrees) of the center (centronum 8

20 ZIPTWO 99 char 5

21 ZIP\_CLASS 2 ZIP Code Classification:P=PO Box U=Uniquchar 1

10 most frequent values of MSA

(Metro Statistical Area code by common po)

338 distinct values in total

Rank Value Frequency

---- ----- ---------

1 0 18,804

2 8840 766

OPTIMUM LENGTHS TO MINIMIZE STORAGE (72 bytes shorter)

VARIABLE\_ NEW\_

Obs \_\_TYP TYPE NAME LENGTH ORIGINAL SAVINGS

1 C Character ALIAS\_CITY 272 300 28

2 C Character ALIAS\_CITYN 300 300 0

3 C Character AREACODES 11 12 1

4 C Character CITY 28 35 7

5 C Character COUNTYNM 25 25 0

6 C Character DST 1 1 0

7 C Character PONAME 28 35 7

8 C Character STATECODE 2 2 0

9 C Character STATENAME 25 25 0

10 C Character TIMEZONE 8 9 1

11 C Character ZIP\_CLASS 1 1 0

12 N Numeric AREACODE 3 8 5

13 N Numeric COUNTY 3 8 5

14 N Numeric GMTOFFSET 3 8 5

15 N Numeric MSA 4 8 4

16 N Numeric STATE 3 8 5

17 N Numeric X 8 8 0

18 N Numeric Y 8 8 0

19 N Numeric ZIP 4 8 4

==

72 Bytes

These char variables have equal values for all observations

BATCH

Variables with All Equal Values

Variable Type Len Compare Len Label Compare Label

CITY CHAR 35 PONAME 35 Name of city/org USPS Post Office Name: same as City

Variable Correlations

Correlated Correlation Number

Variable With Coef of Obs

X ZIP 0.96130 41466 \* spearman longitude and zip;

GMTOFFSET X 0.92613 41466 \* Greenich meantime and zip (like logitude)

GMTOFFSET ZIP 0.88619 41466

Relationship OF VARIABLES WHERE MAX LEVELS IS 400 AND MAX NUMBER OF VARIABLES IS 10

One to One -- One to many -- Many to One -- Many to Many

32 One to One GMTOFFSET to TIMEZONE

46 One to One STATE to STATECODE

47 One to One STATE to STATENAME

13 Many to One CNT to STATE \* I created the cnt variable to test Many to One

14 Many to One CNT to STATECODE

15 Many to One CNT to STATENAME

31 Many to Many GMTOFFSET to STATENAME

Cramers V (like a correlation except with character variables)

‘1’ means probably a 1:1 relationship)

ALL PAIRS OF VARIABLES WHERE MAX LEVELS IS 2000 AND MAX NUMBER OF VARIABLES IS 30

TIMEZONE \*

Obs TABLE STATISTIC VALUE

1 Table STATE \* STATECODE Cramer's V 1

2 Table STATE \* STATENAME Cramer's V 1

3 Table STATE \* STATENAME2 Cramer's V 1

4 Table STATECODE \* STATENAME Cramer's V 1

5 Table STATECODE \* STATENAME2 Cramer's V 1

6 Table STATENAME \* STATENAME2 Cramer's V 1

7 Table GMTOFFSET \* TIMEZONE Cramer's V 1

8 Table AREACODE \* AREACODES Cramer's V 1

9 Table STATE \* AREACODES Cramer's V 0.9951578

10 Table AREACODE \* STATE Cramer's V 0.9951578

11 Table AREACODE \* STATECODE Cramer's V 0.9951578

12 Table AREACODE \* STATENAME Cramer's V 0.9951578

13 Table AREACODE \* STATENAME2 Cramer's V 0.9951578

14 Table AREACODES \* STATENAME Cramer's V 0.9951578

15 Table AREACODES \* STATENAME2 Cramer's V 0.9951578

16 Table AREACODES \* STATECODE Cramer's V 0.9951578

The data analyzer is capable of finding all two part primary keys

For instance the combination of latitude and longitude do not form

A primary key

The primary key needs to have a cardinality of 41,267 like ZIP

Lat and long combination do nto form a primary key

Obs RECORDS VAR1ST LEVELS\_1 VAR2ND LEVELS\_2 UNIQUE

48 41,267 Y 41,158 X 41,207 41,233

May be due to 0,0 combinations?

10 most frequent values of Y

(Latitude (degrees) of the center (centro)

41,158 distinct values in total

Rank Value Frequency

---- ----- ---------

1 38.899101 8

2 0.000000 5

3 36.171909 4

4 32.514610 3

10 most frequent values of X

(Longitude (degrees) of the center (centr)

41,207 distinct values in total

Rank Value Frequency

---- ----- ---------

1 -77.028999 8

2 0.000000 5

3 -115.139969 4

4 -122.441650 3

5 -85.766403 3

VooDoo does missing pattern analysis

DATA VERIFICATION + VALIDATION FOR WORK.ZIPCODE

zipcode

Variable Name Aliases for \_vvcop for Use in Mising Value Pattern Analysis

Sorted by ALIAS

F

A

K

E

\_

F M

F A I

F A K S

F A K E S

A K E \_ W

K E \_ S I A

E \_ A O T A L S

\_ S L M H Z S S A G L I T

O A L E C I T T C A R T M I A A

P N M M M O P A A O R E I T A S T M

E E E I I N \_ T T C U E A M O P S \_ E I

R G V V S S S C S E E O N A C E F O \_ C C N S

F C R A D A S S T L C T C N U T C O Z F N C I I A S

O R E O L A L I I A Z A I A O A N Y M O D O S D A I T T M I

b E N U U T U N N N I S T T D M T N S D E N E S M T Y Y E N

s Q T P E E E G G T P Y X S Y E E E Y M A E S E T T E Y N 2 2 G

1 10061 24.4 24 X X X . . . X X X . X X X X X X X X X X X X X . . X X 6

2 9175 22.2 16 X X X . X . X X X . X X X X X X X X X X X X X . . X X 5

3 5100 12.4 20 X X X . . . X X X X X X X X X X X X X X X X X . . X X 5

4 4006 9.7 12 X X X . X . X X X X X X X X X X X X X X X X X . . X X 4

5 3476 8.4 22 X X X . . . X X X . X X X X X X X X X X X X X X . X X 5

6 2523 6.1 14 X X X . X . X X X . X X X X X X X X X X X X X X . X X 4

7 1496 3.6 23 X X X . . . X X X . X X X X X X X X X X X X X . X X X 5

8 1111 2.7 15 X X X . X . X X X . X X X X X X X X X X X X X . X X X 4

9 819 2.0 18 X X X . . . X X X X X X X X X X X X X X X X X X . X X 4

10 687 1.7 21 X X X . . . X X X . X X X X X X X X X X X X X X X X X 4

11 571 1.4 8 X X X . X X X X X . X X X X X X X X X X X X X . . X X 4

12 543 1.3 10 X X X . X . X X X X X X X X X X X X X X X X X X . X X 3

13 452 1.1 13 X X X . X . X X X . X X X X X X X X X X X X X X X X X 3

14 343 0.8 4 X X X . X X X X X X X X X X X X X X X X X X X . . X X 3

15 309 0.7 19 X X X . . . X X X X X X X X X X X X X X X X X . X X X 4

16 153 0.4 11 X X X . X . X X X X X X X X X X X X X X X X X . X X X 3

17 140 0.3 6 X X X . X X X X X . X X X X X X X X X X X X X X . X X 3

18 93 0.2 7 X X X . X X X X X . X X X X X X X X X X X X X . X X X 3

19 80 0.2 2 X X X . X X X X X X X X X X X X X X X X X X X X . X X 2

20 52 0.1 17 X X X . . . X X X X X X X X X X X X X X X X X X X X X 3

21 37 0.1 9 X X X . X . X X X X X X X X X X X X X X X X X X X X X 2

22 27 0.1 5 X X X . X X X X X . X X X X X X X X X X X X X X X X X 2

23 8 0.0 1 X X X . X X X X X X X X X X X X X X X X X X X X X X X 1

24 5 0.0 3 X X X . X X X X X X X X X X X X X X X X X X X . X X X 2

If you are going to to multiple imputation then you will want to know which variables are monotonic.

These two columns display

the montotonic missing pattern.

If fake some missing is missing

then fake missing with constant is missing.

Monotonic missing are easier to

impute than other patterns.

f

a

k

e

\_

f m

a i

k s

e s

\_ w

s i

o t

m h

e c

m o

i n

s s

s t

i a

n n

g t

. .

. .

. .

. .

. .

. .

. .

. .

X .

X .

X .

X .

X .

X .

X .

X .

X X

X X

X X

X X

X X

X X

X X

X X