CHICAGO- DALLAS FOOD INSPECTION

Data Type Adjustments :

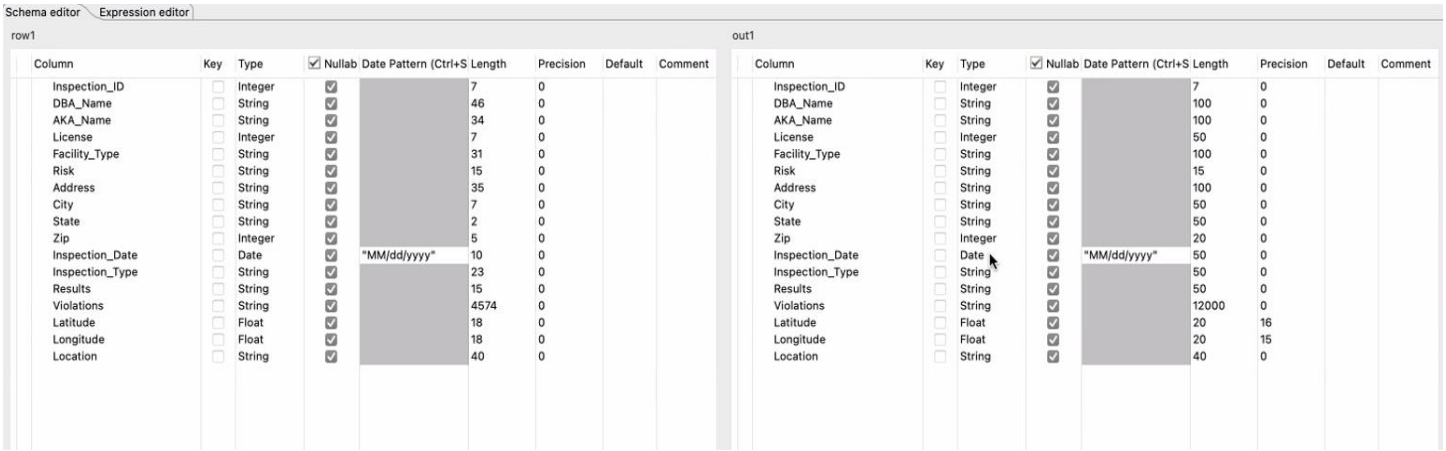
In the Chicago Food Inspection dataset all the columns (variables) were in string data type which not proper/desirable for analysis. We had to adjust the data types before we could start working.
Following are the adjustments made :

- **Inspection date** : The default data type of the dataset was set to **string**; it was converted to **date** data type. The date data type automatically accounted for both date and time in Talend environment.
- **Latitude :** This column was converted to **float** from number data type. This was done to get proper geographical representation while plotting geographical type visualization
- **Longitude :** Similar adjustments were performed on this column as well as done to the latitude column.
- **Zip code :** This column was string as well by default while loading . It was later converted to integer
- **Dealing with "violations" column**:
  The data type for this column was string by default and as it as description-based column, we kept the data type as it , but , had to work on other challenges while working on this column.

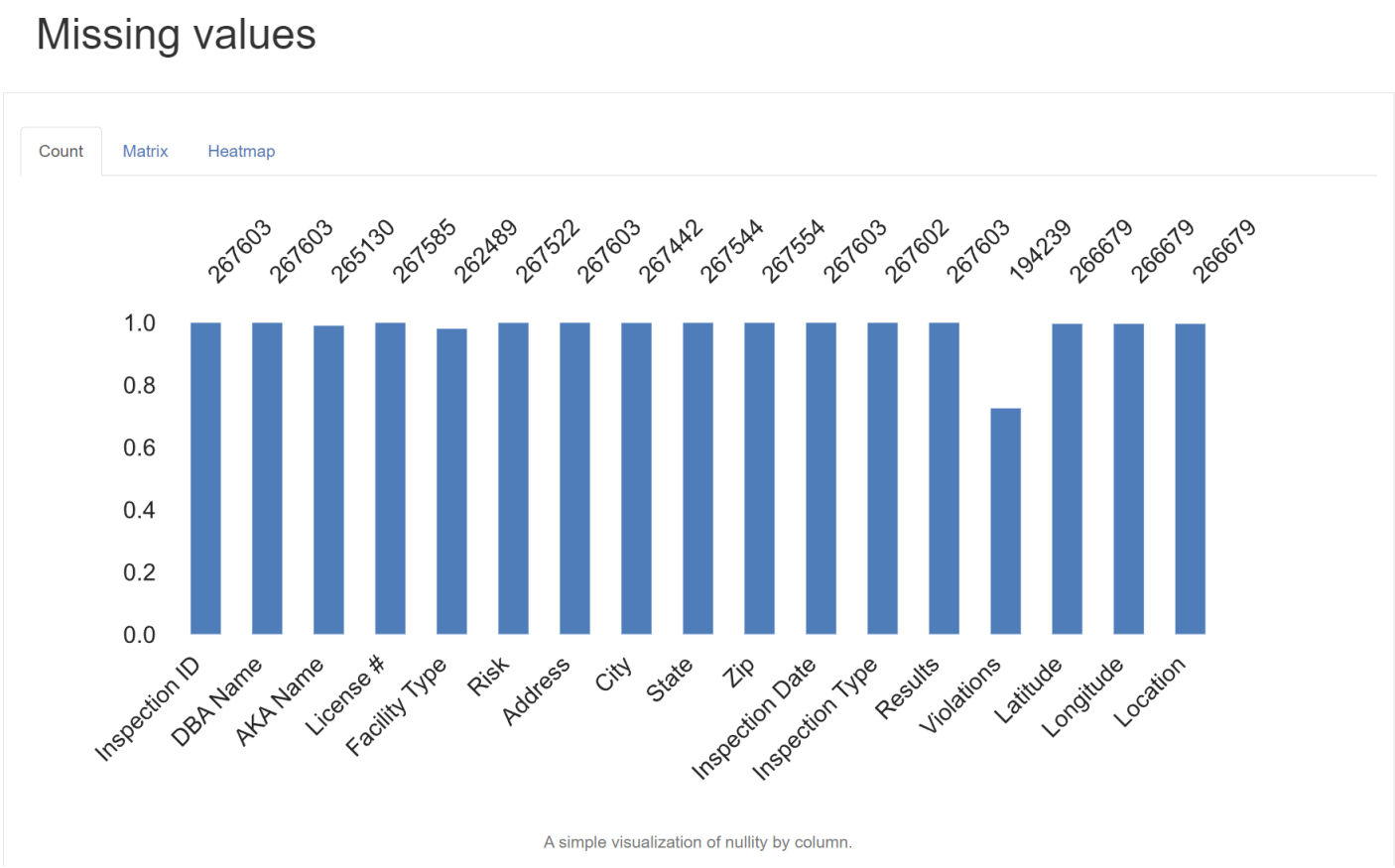
By default, in the Chicago food inspection dataset, multiples violations were entered in a single row. That is, given a restaurant with multiple violations, all of them are present in a single row. This proves to be inconvenient while analyzing each inspection ID.
Furthermore, each violation column consists of 3 main parts : violation code, violation description and comment.
To deal with this, we needed to normalize the violation column using a component known as t-normalise in Talend. As a result, each violation is split around the '|' operator leading to each violation occupying a single row. At this point, each column consists of 3 parts as mentioned above : violation code, violation description and comment.
Now, in order to split each part into separate columns to clearly analyze the data, we used a regex function called : "extract regex field". This helped in splitting each of these sections to be split into separate columns.

COLUMN SIZE ADJUSTMENTS :

The default size allocations of each column were improper and acted as an hinderance while working the dataset. There were issues while staging the data in the SQL database. So we had to adjust the column sizes manually ,so as to provide proper memory allocation thus leading to proper staging of the data in the data bases



MISSING VALUES COLUMN WISE :



COLUMN WISE ANALYSIS :

1. **Inspection ID** : All the values are unique with a maximum value of 2589562 and minimum value of 44247. There are no missing values in this column.

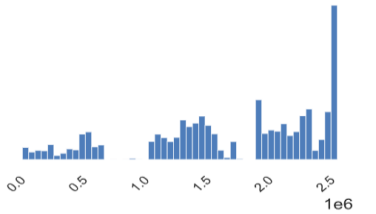
Inspection ID

Inspection ID

Real number (ℝ)

UNIQUE

Distinct	267603	**Minimum**	44247
Distinct (%)	100.0%	**Maximum**	2589562
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1729946.1	**Memory size**	2.0 MiB



More details

Statistics | Histogram | Common values | Extreme values

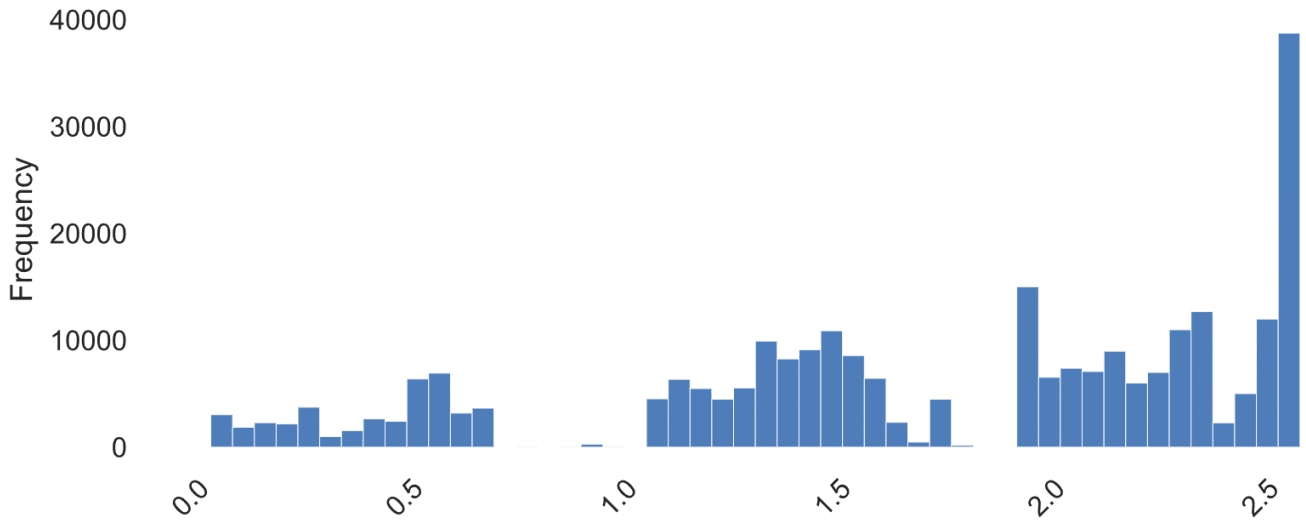
Quantile statistics

Minimum	44247
5-th percentile	327242.1
Q1	1307369.5
median	1950867
Q3	2357001.5
95-th percentile	2575762.9
Maximum	2589562
Range	2545315
Interquartile range (IQR)	1049632

Descriptive statistics

Standard deviation	719474.86
Coefficient of variation (CV)	0.41589438
Kurtosis	-0.65210317
Mean	1729946.1
Median Absolute Deviation (MAD)	547189
Skewness	-0.62804959
Sum	4.6293877×10^{11}
Variance	5.1764407×10^{11}
Monotonicity	Not monotonic

Values distributions by percentage :



2.**DBA NAME :** This is a descriptive column with 32170 unique values, that is up to 12% of the total number of the total. The longest dba name being of 79 characters and shortest being 49 characters.

DBA Name

Text

Distinct	32170
Distinct (%)	12.0%
Missing	0
Missing (%)	0.0%
Memory size	2.0 MiB

More details

Overview | Words | Characters

Length

Max length	79
Median length	49
Mean length	18.731875
Min length	1

Characters and Unicode

Total characters	5012706
Distinct characters	83
Distinct categories	12
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	5939
Unique (%)	2.2%

Sample

1st row	SUBWAY SANDWICH & SALAD
2nd row	GEORGE E TAYLOR CHILD DEVELOPMENT CENTER
3rd row	CASA CENTRAL COMMUNITY SERVICE
4th row	Drummond
5th row	STICKY KISS

3. **AKA Name** : This column has only 30619 distinct values, that is approximately 11.5% of the total. It has total of 2473 missing values (0.9%). The longest dba name being of 79 characters and shortest being 49 characters.

AKA Name

Text

Distinct	30619
Distinct (%)	11.5%
Missing	2473
Missing (%)	0.9%
Memory size	2.0 MiB

More details

Overview | Words | Characters

Length

Max length	79
Median length	49
Mean length	17.914834
Min length	2

Characters and Unicode

Total characters	4749760
Distinct characters	87
Distinct categories	13
Distinct scripts	2
Distinct blocks	2

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	5781
Unique (%)	2.2%

Sample

1st row	SUBWAY SANDWICH & SALAD
2nd row	GEORGE E TAYLOR CHILD DEVELOPMENT CENTER
3rd row	CASA CENTRAL COMMUNITY SERVICE
4th row	Drummond
5th row	STICKY KISS

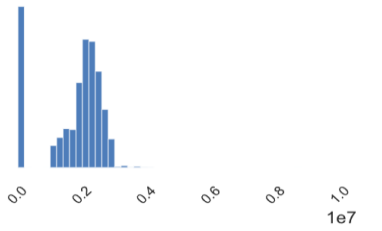
4. License# : This column has total 16.6% distinct values(44510). With a minimum value being 0 and largest value being 9999999. It has only 18 missing values i.e. <0.1% of the total, with a maximum length 47 and minimum length of 10

License

Real number (ℝ)

Distinct	44510
Distinct (%)	16.6%
Missing	18
Missing (%)	< 0.1%
Infinite	0
Infinite (%)	0.0%
Mean	1721777

Minimum	0
Maximum	9999999
Zeros	691
Zeros (%)	0.3%
Negative	0
Negative (%)	0.0%
Memory size	2.0 MiB



More details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	0
5-th percentile	17602
Q1	1336796
median	2048785
Q3	2369187
95-th percentile	2766632
Maximum	9999999
Range	9999999
Interquartile range (IQR)	1032391

Descriptive statistics

Standard deviation	927007.62
Coefficient of variation (CV)	0.53840167
Kurtosis	-0.42600222
Mean	1721777
Median Absolute Deviation (MAD)	420444
Skewness	-0.8968191
Sum	4.6072171×10^{11}
Variance	8.5934313×10^{11}
Monotonicity	Not monotonic

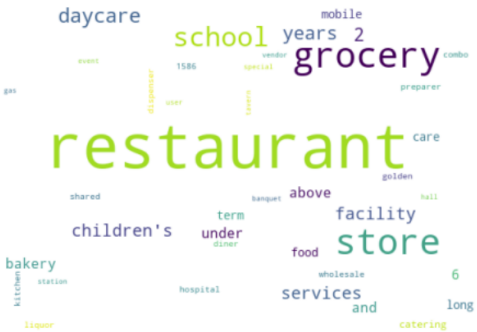
5. FACILITY TYPE : This column has only 513 values, that is approximately 0.2% of the total values. It has approximately 5114 missing values (1.9%) and 73 distinct values (<0.1%)

Facility Type

Text

MISSING

Distinct	513
Distinct (%)	0.2%
Missing	5114
Missing (%)	1.9%
Memory size	2.0 MiB



More details

Overview | Words | Characters

Length

Max length	47
Median length	10
Mean length	11.105433
Min length	3

Characters and Unicode

Total characters	2915054
Distinct characters	69
Distinct categories	9
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	73
Unique (%)	< 0.1%

Sample

1st row	Restaurant
2nd row	Daycare (2 - 6 Years)
3rd row	Restaurant
4th row	School
5th row	Restaurant

6.RISK : This column has 4 distinct values and 81 missing values both being (<0.1%). This columns has max length of 15 characters, minimum length of 3 characters and median length of 13 characters.

Risk

Categorical

Distinct	4
Distinct (%)	< 0.1%
Missing	81
Missing (%)	< 0.1%
Memory size	2.0 MiB

More details

Overview | Categories | Words | Characters

Length

Max length	15
Median length	13
Mean length	13.287109
Min length	3

Characters and Unicode

Total characters	3554594
Distinct characters	23
Distinct categories	6
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Risk 1 (High)
2nd row	Risk 1 (High)
3rd row	Risk 1 (High)
4th row	Risk 1 (High)
5th row	Risk 2 (Medium)

7. Address : This is a text-based column with 19641 distinct values i.e. 7.3% of the total. Max length of the column is 52 and min length is 1 and median length. This column has 2049 distinct values i.e. 0.8% value.

Address

Text

Distinct	19641
Distinct (%)	7.3%
Missing	0
Missing (%)	0.0%
Memory size	2.0 MiB

More details

Overview | Words | Characters

Length

Max length	52
Median length	40
Mean length	18.610143
Min length	1

Characters and Unicode

Total characters	4980130
Distinct characters	72
Distinct categories	10
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	2049
Unique (%)	0.8%

Sample

1st row	2352 W CERMAK RD
2nd row	3232 W DIVISION ST
3rd row	1343 N CALIFORNIA AVE
4th row	1845 W Cortland ST
5th row	925 N ASHLAND AVE

8. **City :** This column has string data type with 79 distinct values. min length of 7, median length = 7 and max of 20 characters. There are a total of 34 unique values.

City

City
Text

Distinct	79
Distinct (%)	< 0.1%
Missing	161
Missing (%)	0.1%
Memory size	2.0 MiB

More details

Overview | Words | Characters

Length

Max length	20
Median length	7
Mean length	7.0028268
Min length	2

Characters and Unicode

Total characters	1872850
Distinct characters	44
Distinct categories	5
Distinct scripts	2
Distinct blocks	1

Unique

Unique	34
Unique (%)	< 0.1%

Sample

1st row	CHICAGO
2nd row	CHICAGO
3rd row	CHICAGO
4th row	CHICAGO
5th row	CHICAGO

9. **State :** This column has a high correlation and is imbalanced. It has 5 distinct values and 59 missing values. This column has same values for max, median and min length, i.e. 2.

State

State

Categorical

HIGH CORRELATION IMBALANCE

Distinct	5
Distinct (%)	< 0.1%
Missing	59
Missing (%)	< 0.1%
Memory size	2.0 MiB

More details

Overview | Categories | Words | Characters

Length

Max length	2
Median length	2
Mean length	2
Min length	2

Characters and Unicode

Total characters	535088
Distinct characters	7
Distinct categories	1
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	2
Unique (%)	< 0.1%

Sample

1st row	IL
2nd row	IL
3rd row	IL
4th row	IL
5th row	IL

10. **ZIP :** This column has 121 distinct values and 49 missing values. The minimum value for this column is : 10014 and maximum value is : 90504. The data of this column is skewed in nature.

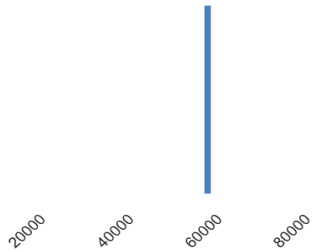
Zip

Real number (ℝ)

SKEWED

Distinct	121
Distinct (%)	< 0.1%
Missing	49
Missing (%)	< 0.1%
Infinite	0
Infinite (%)	0.0%
Mean	60628.705

Minimum	10014
Maximum	90504
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	2.0 MiB



More details

Statistics | Histogram | Common values | Extreme values

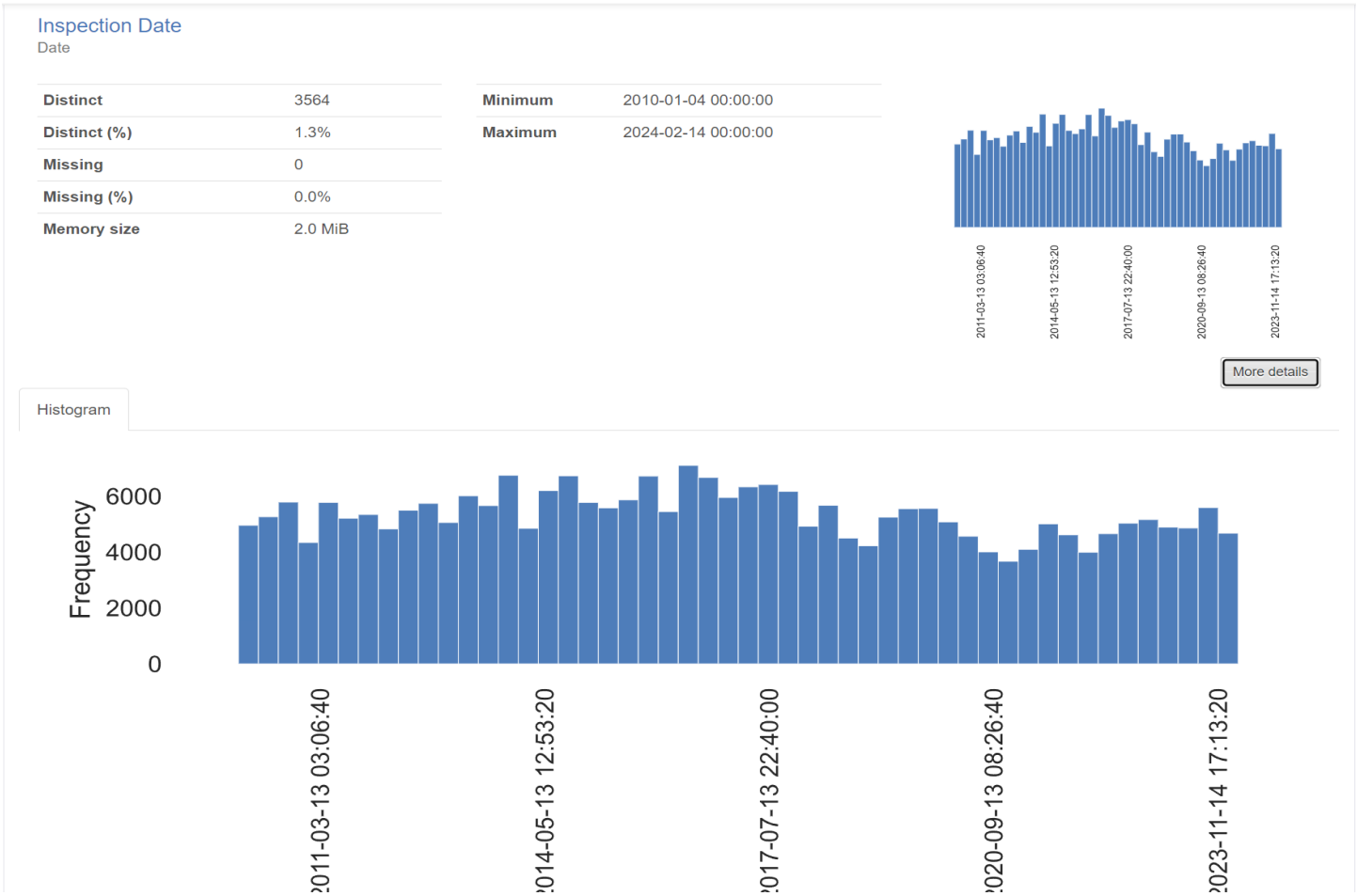
Quantile statistics

Minimum	10014
5-th percentile	60605
Q1	60614
median	60625
Q3	60643
95-th percentile	60659
Maximum	90504
Range	80490
Interquartile range (IQR)	29

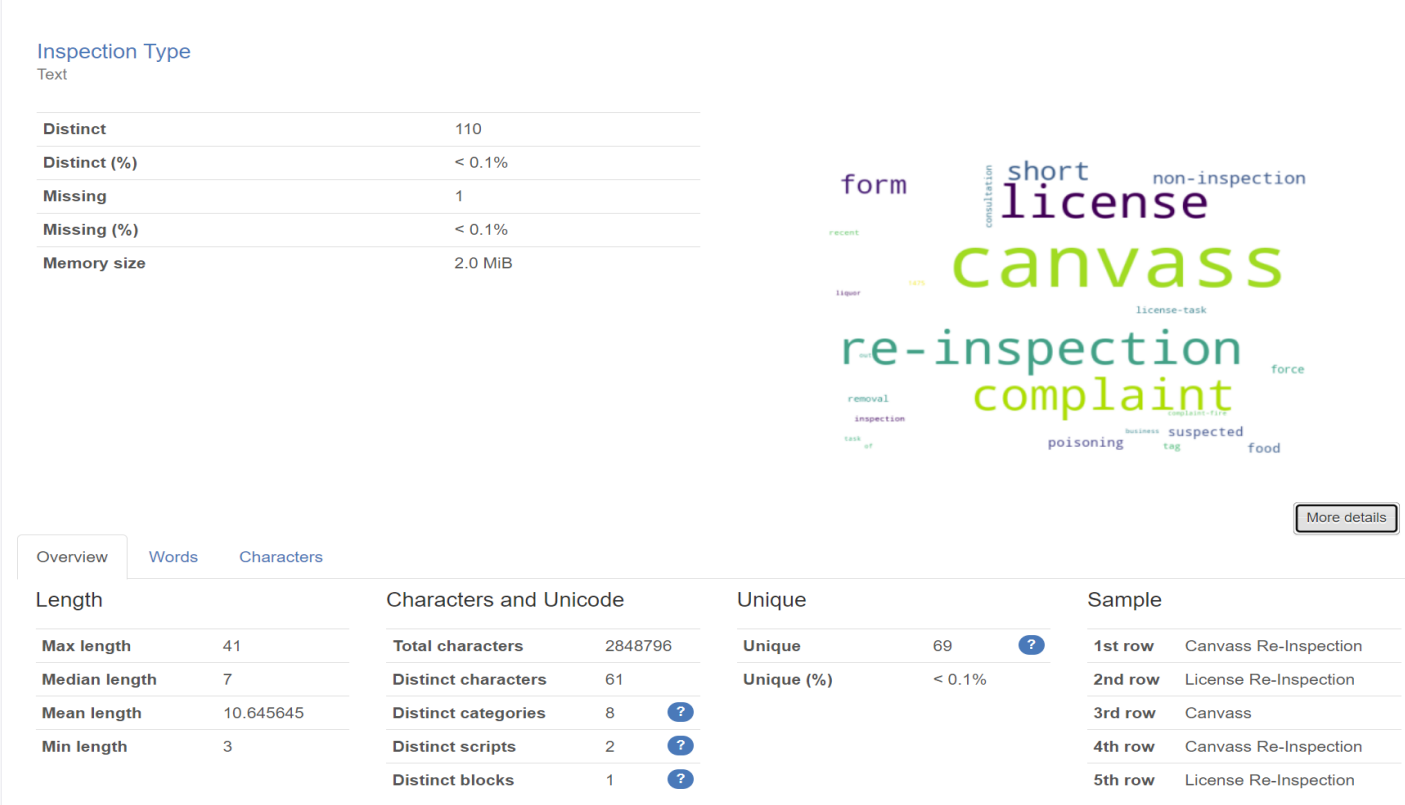
Descriptive statistics

Standard deviation	148.86011
Coefficient of variation (CV)	0.0024552745
Kurtosis	63946.227
Mean	60628.705
Median Absolute Deviation (MAD)	14
Skewness	-111.18274
Sum	1.6221452×10^{10}
Variance	22159.333
Monotonicity	Not monotonic

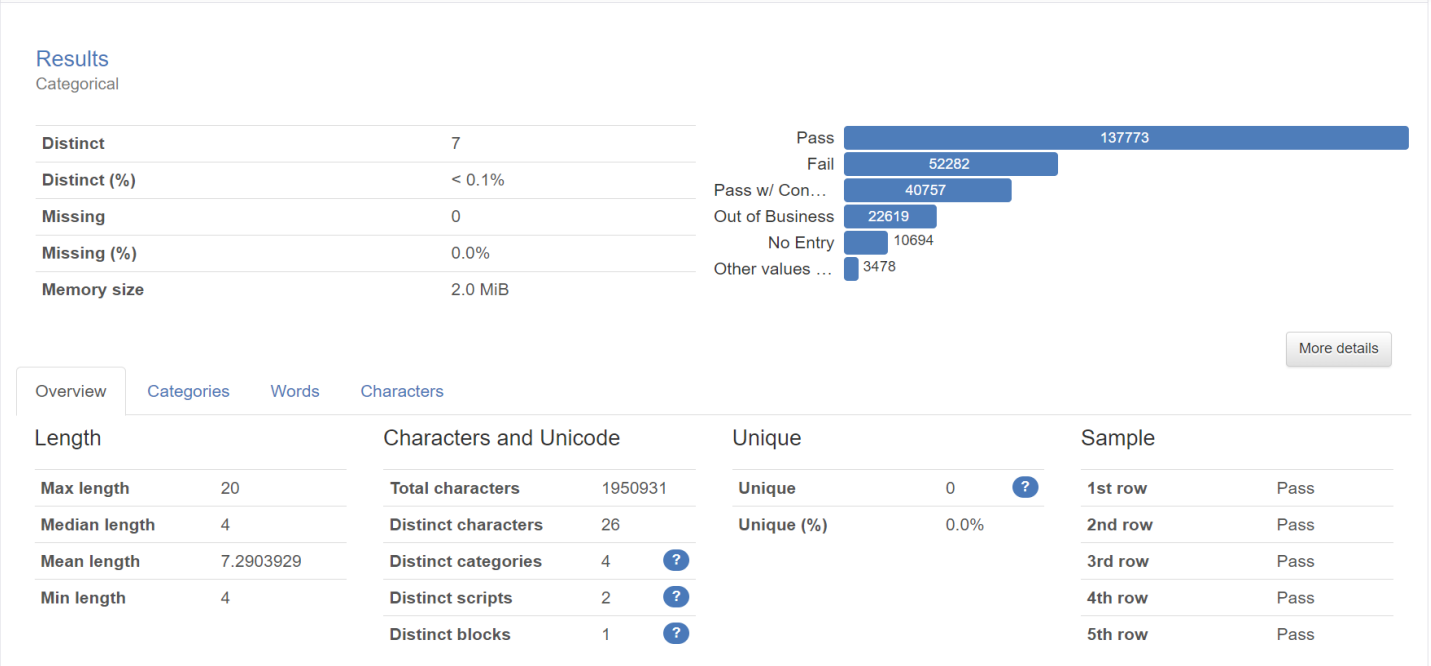
11. **Inspection Date :** Despite the data type of this column being date, it accounts for both date and time. This column has 3564 distinct i.e. (around 1.3%) minimum : 2010-01-04 00:00:00 and max : 2024-02-14 00:00:00



12. **Inspection Type :** This column has 110 distinct values that accounts for 0.1% of the total number of rows. There is only one missing value. The content of the column varies from min of 3 characters to max of 41 having 7 as median.



13. Results : The results column gives about the result of the inspection. It has 7 distinct values as shown in the diagram. It has no missing values.



14. Violations : Most of the analysis revolves around this column. The data present in this column was in a very raw and unprocessed state. It has 192790 distinct values(99.3%) and 73364 missing values (27.4%). The max length of content of this column is 11620 characters and min length is 30 characters.

Violations
Text

MISSING

Distinct	192970
Distinct (%)	99.3%
Missing	73364
Missing (%)	27.4%
Memory size	2.0 MiB

More details

Overview | Words | Characters

Length

Max length	11620
Median length	5021
Mean length	1221.1861
Min length	30

Characters and Unicode

Total characters	237201973
Distinct characters	100
Distinct categories	14
Distinct scripts	2
Distinct blocks	2

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	192053
Unique (%)	98.9%

Sample

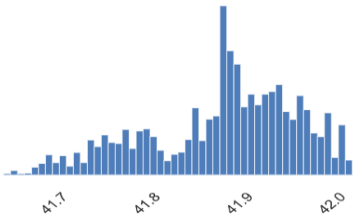
1st row	10. ADEQUATE HANDWASHING SINKS PROPERLY SUPPLIED AND ACCESSIBLE - Comments: 6-301.14 OBSERVED NO HAND WASHING SIGN AT THE REAR HAND WASHING SINK. INSTRUCTED TO PROVIDE. | 58. ALLERGEN TRAINING AS REQUIRED - Comments: 2-102.13 OBSERVED THE MANAGER ON DUTY WITHOUT STATE ALERGENT CERTIFICATE. INSTRUCTED

15. Latitude : This column has float type data type and consists of 18211 distinct values and 924 missing rows.

Latitude
Real number (ℝ)

HIGH CORRELATION

Distinct	18211	Minimum	41.64467
Distinct (%)	6.8%	Maximum	42.021064
Missing	924	Zeros	0
Missing (%)	0.3%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	41.880731	Memory size	2.0 MiB



More details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	41.64467
5-th percentile	41.726917
Q1	41.831831
median	41.89177
Q3	41.939785
95-th percentile	41.997516
Maximum	42.021064
Range	0.37639412
Interquartile range (IQR)	0.10795422

Descriptive statistics

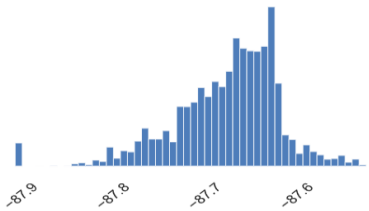
Standard deviation	0.081070003
Coefficient of variation (CV)	0.0019357351
Kurtosis	-0.35408336
Mean	41.880731
Median Absolute Deviation (MAD)	0.050608465
Skewness	-0.58981541
Sum	11168711
Variance	0.0065723453
Monotonicity	Not monotonic

16. Longitude : This column has float type data type and consists of 18211 distinct values and 924 missing rows.

Longitude
Real number (ℝ)

HIGH CORRELATION

Distinct	18211	Minimum	-87.906874
Distinct (%)	6.8%	Maximum	-87.525094
Missing	924	Zeros	0
Missing (%)	0.3%	Zeros (%)	0.0%
Infinite	0	Negative	266679
Infinite (%)	0.0%	Negative (%)	99.7%
Mean	-87.676339	Memory size	2.0 MiB



More details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	-87.906874
5-th percentile	-87.777776
Q1	-87.707517
median	-87.666419
Q3	-87.634942
95-th percentile	-87.59954
Maximum	-87.525094
Range	0.38178
Interquartile range (IQR)	0.072575797

Descriptive statistics

Standard deviation	0.058260207
Coefficient of variation (CV)	-0.00066449178
Kurtosis	2.0976425
Mean	-87.676339
Median Absolute Deviation (MAD)	0.034445649
Skewness	-0.97642782
Sum	-23381438
Variance	0.0033942517
Monotonicity	Not monotonic

Alteryx analysis :

1 of 1 Fields | Records 1 to 1

Record	Report
1	**String/Character Fields**

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
AKA Name	0.9%	30,628	BP	MAE'S EARLY CHILDHOOD DEVELOPMENT AND THERAPEUTIC DAY CARE CENTER, INCORPORATED	1	4,376	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Longitude	0.3%	18,216	-87.669464645	-87.69538780161594	1	3,291	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Address	0.0%	19,645		1400 S JEAN BAPTISTE POINTE DUSABLE LAKESHORE DRIVE	1	3,273	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
State	0.0%	6	IL	IL	1	267,681	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
DBA Name	0.0%	32,180	N	MAE'S EARLY CHILDHOOD DEVELOPMENT AND THERAPEUTIC DAY CARE CENTER, INCORPORATED	1	3,556	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Inspection ID	0.0%	267,751	58558	2589679	1	1	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
License #	0.0%	44,527	0	2948383	1	692	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Latitude	0.3%	18,216	41.786130727	41.799368015121125	1	3,291	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
City	0.1%	80	CH	BANNOCKBURNDEERFIELD	1	266,689	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Violations	27.4%	166,199	63. REMOVAL OF SUSPENSION SIGN	2. CITY OF CHICAGO FOOD SERVICE SANITATION CERTIFICATE - Comments: OBSERVED FACILITY FOOD HANDLERS PREPARING AND HANDLING ï¿½TIME AND TEMPERATURE CONTROL FOR SAFETY(TCS) FOODS WITHOUT ORIGINAL CITY OF CHICAGO FOOD SERVICE MANAGER AND CERTIFICATE ON SITE.	1	73,407	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Risk	0.0%	5	All	Risk 2 (Medium)	56	196,229	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Location	0.3%	18,216	(41.9960984214, -87.7868769104594)	(41.799368015121125, -87.59468625684886)	1	3,291	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Inspection Type	0.0%	111	SFP	LICENSE TASK FORCE / NOT -FOR-PROFIT CLUB	1	139,149	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Inspection Date	0.0%	3,566	02/15/2024	02/15/2024	1	185	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Facility Type	1.9%	514	bar	MOBILE FROZEN DESSERTS DISPENSER-NON-MOTORIZED	1	179,935	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Zip	0.0%	122	60608	60608	1	9,863	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Results	0.0%	7	Pass	Business Not Located	86	137,862	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.

DALAS FOOD INSPECTION :

CORRELATIONS :

Heatmap | Table

	Inspection Score	Inspection Type	Inspection Year	Street Direction	Street Number	Street Type
Inspection Score	1.000	0.088	0.026	0.043	0.007	0.032
Inspection Type	0.088	1.000	0.025	0.010	-0.013	0.023
Inspection Year	0.026	0.025	1.000	0.033	0.005	0.020
Street Direction	0.043	0.010	0.033	1.000	-0.166	0.450
Street Number	0.007	-0.013	0.005	-0.166	1.000	0.258
Street Type	0.032	0.023	0.020	0.450	0.258	1.000

Among all the columns, correlations among these columns are more crucial. Other columns not shown here are mainly the violations, comments columns which are cleaned and analyzed in detail. The latitude and longitude columns are also highly corelated.

1 Restaurant Name : This column has string data type. It has 9136 distinct values (11.7%) and 11 missing rows. Also the name sizes vary from min of 3 to a max length of 65 characters having 52 characters as median.

Restaurant Name
Text

Distinct	9136
Distinct (%)	11.7%
Missing	11
Missing (%)	< 0.1%
Memory size	612.6 KiB

More details

Overview | Words | Characters

Length

Max length	65
Median length	52
Mean length	19.248402
Min length	3

Characters and Unicode

Total characters	1508863
Distinct characters	77
Distinct categories	12
Distinct scripts	2
Distinct blocks	2

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	854
Unique (%)	1.1%

Sample

1st row	MICKLE CHICKEN
2nd row	TOM THUMB - JUICE BAR
3rd row	BROOKDALE WHITE ROCK
4th row	CHURCH'S CHICKEN #201
5th row	PEAK PREPARATORY-PRIMARY SCHOOL

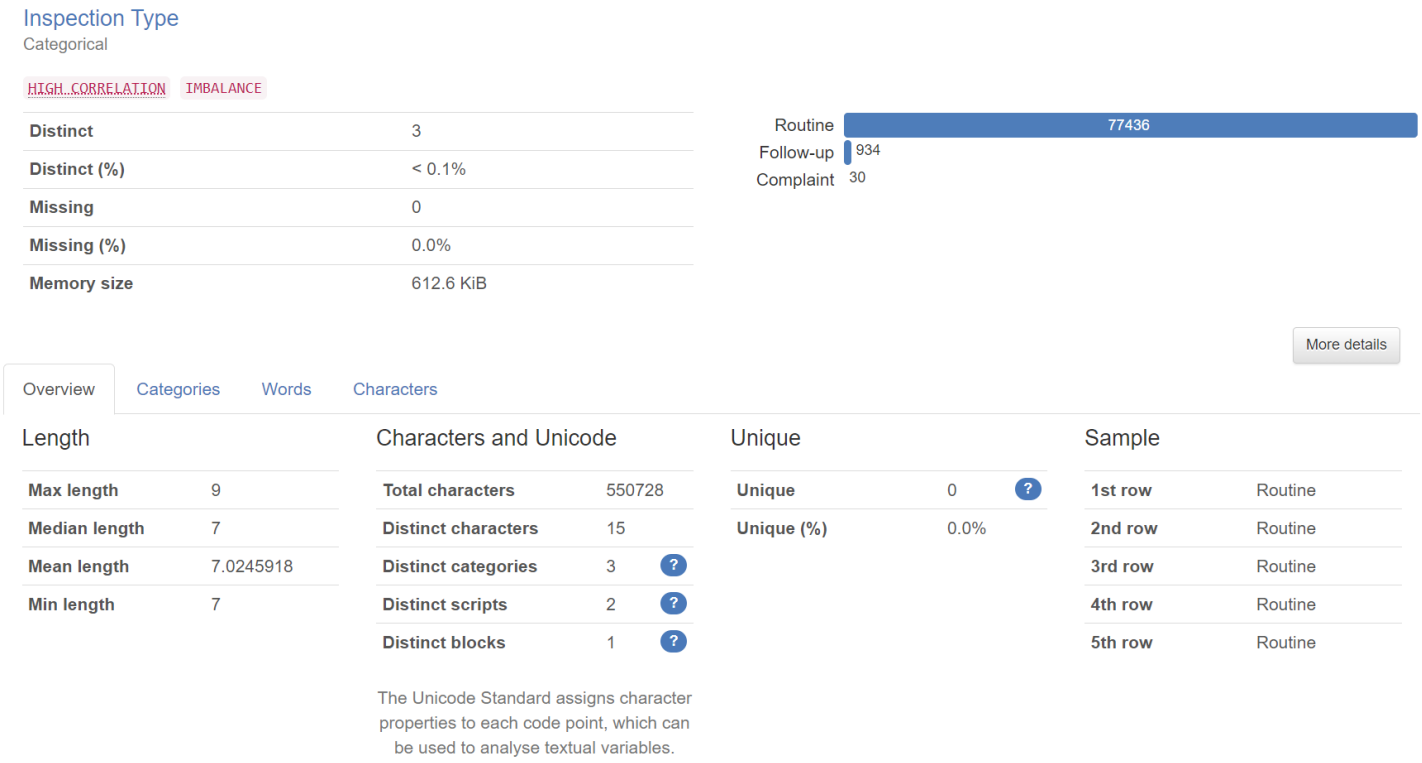
Here is a list of most commonly used words present in the Restaurant name column :

Overview | Words | Characters

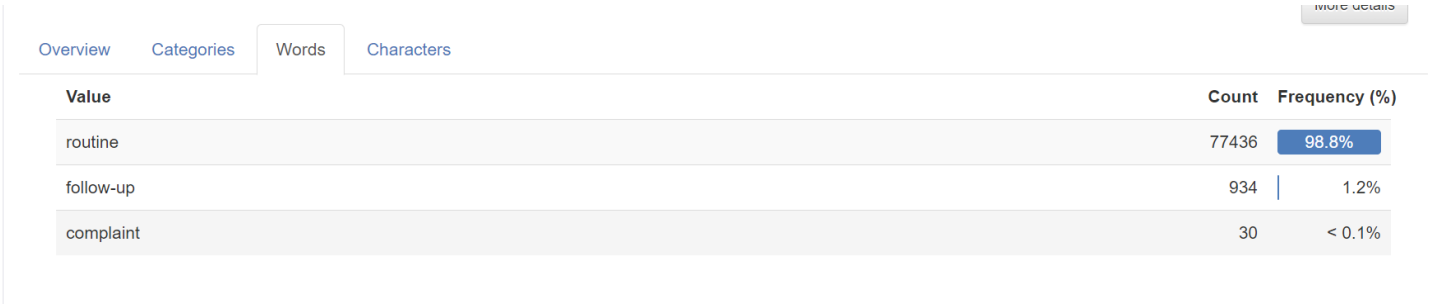
Value	Count	Frequency (%)
	7423	3.0%
bar	3176	1.3%
restaurant	2873	1.2%
the	2803	1.1%
food	2714	1.1%
kitchen	2645	1.1%
cafe	2563	1.0%
la	2488	1.0%
el	2276	0.9%
school	2167	0.9%
Other values (7446)	214728	87.3%

el

2. Inspection Type : This column has a high correlation and is imbalanced. It has 3 distinct values : Routine, Follow-up and Complaint. It has no missing values.



Here is the percentage wise occurrence of each category of inspection type.

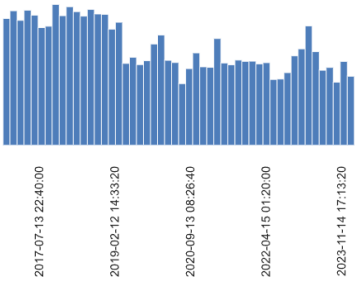


3 INSPECTION DATE : This column states the date on which the inspection was conducted on the restaurant. It has 2293 distinct values(2.9%) and no missing values whatsoever. There is a graph showing the distribution of count on how many inspections were conducted each

Inspection Date
Date

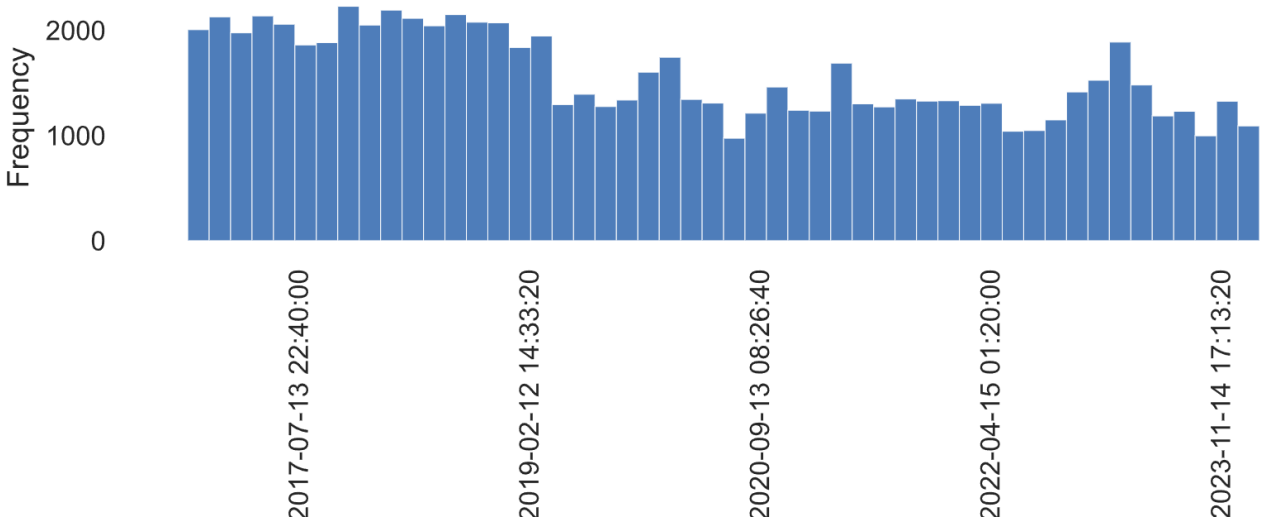
Distinct	2293
Distinct (%)	2.9%
Missing	0
Missing (%)	0.0%
Memory size	612.6 KiB

Minimum	2016-10-03 00:00:00
Maximum	2024-02-14 00:00:00



More details

Histogram



4. INSPECTION SCORE : This column shows the score received by the score received by each restaurant. It has 58 distinct values that is approx.0.1% of the total rows and has no missing values. The highest inspection score is 100 and the minimum inspection score is -26 with a mean of 90.867. Here are the most frequent scores :

Statistics | Histogram | Common values | Extreme values

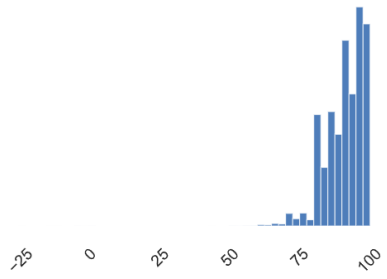
Value	Count	Frequency (%)
100	6525	8.3%
97	5219	6.7%
95	4926	6.3%
94	4696	6.0%
96	4645	5.9%
90	4384	5.6%
92	4228	5.4%
93	4221	5.4%
98	4002	5.1%
91	3921	5.0%
Other values (48)	31633	40.3%

90

Inspection Score

Real number (ℝ)

Distinct	58	**Minimum**	-26
Distinct (%)	0.1%	**Maximum**	100
Missing	0	**Zeros**	2
Missing (%)	0.0%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	6
Infinite (%)	0.0%	**Negative (%)**	< 0.1%
Mean	90.867985	**Memory size**	612.6 KiB



More details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	-26
5-th percentile	80
Q1	87
median	92
Q3	96
95-th percentile	100
Maximum	100
Range	126
Interquartile range (IQR)	9

Descriptive statistics

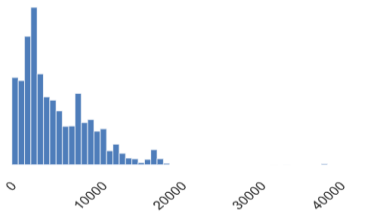
Standard deviation	6.9800247
Coefficient of variation (CV)	0.076815005
Kurtosis	5.0197807
Mean	90.867985
Median Absolute Deviation (MAD)	5
Skewness	-1.1533272
Sum	7124050
Variance	48.720744
Monotonicity	Not monotonic

5. STREET NUMBER : This column is the street address of each restaurant present in the dataset. It has 3442 distinct values, that is 4.4% and no missing values whatsoever. There are 71 zero value rows in this column.

Street Number

Real number (ℝ)

Distinct	3442	**Minimum**	0
Distinct (%)	4.4%	**Maximum**	39779
Missing	0	**Zeros**	71
Missing (%)	0.0%	**Zeros (%)**	0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	5511.2565	**Memory size**	612.6 KiB



More details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	0
5-th percentile	515
Q1	2301
median	4141
Q3	8233
95-th percentile	13434
Maximum	39779
Range	39779
Interquartile range (IQR)	5932

Descriptive statistics

Standard deviation	4333.0204
Coefficient of variation (CV)	0.78621281
Kurtosis	3.6769624
Mean	5511.2565
Median Absolute Deviation (MAD)	2465
Skewness	1.3476421
Sum	4.3208251×10^8
Variance	18775066
Monotonicity	Not monotonic

6. **STREET NAME :** This column has 842 distinct rows that is 1.1% of the total and zero missing values. It has a max length of 25 characters, min of 2 length and median of 22 characters. Also there is the list of freq. used words in the column .

Street Name
Text

Distinct	842
Distinct (%)	1.1%
Missing	0
Missing (%)	0.0%
Memory size	612.6 KiB

More details

Overview | Words | Characters

Length

Max length	25
Median length	22
Mean length	7.8759694
Min length	2

Characters and Unicode

Total characters	617476
Distinct characters	42
Distinct categories	7
Distinct scripts	2
Distinct blocks	1

Unique

Unique	27
Unique (%)	< 0.1%

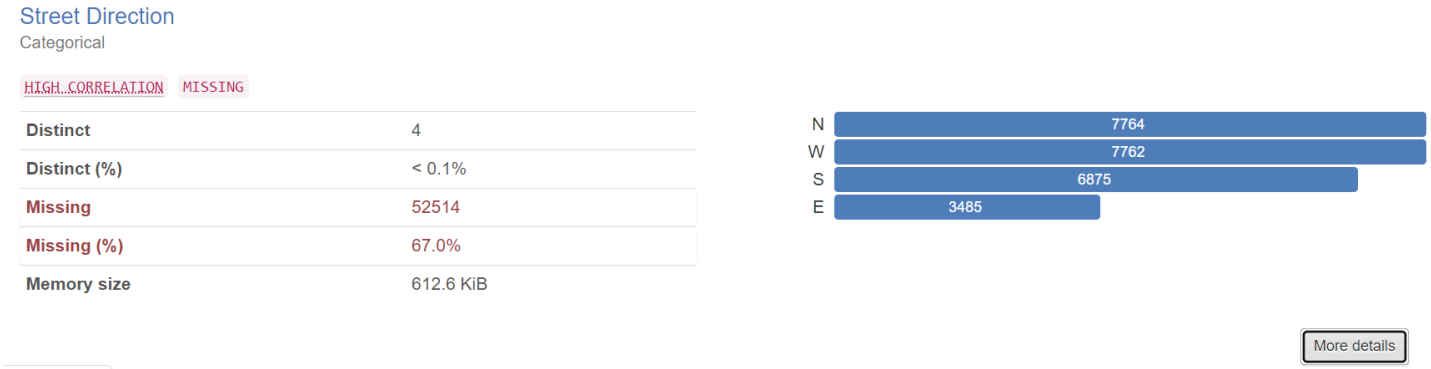
Sample

1st row	CAMP WISDOM
2nd row	FIELD
3rd row	WHITE ROCK
4th row	FERGUSON
5th row	ANNEX

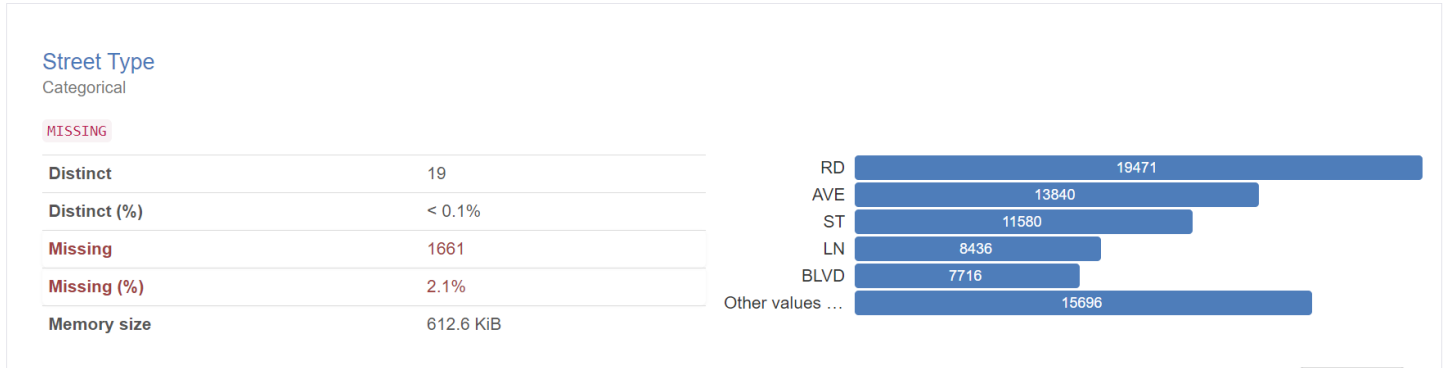
Overview | Words | Characters

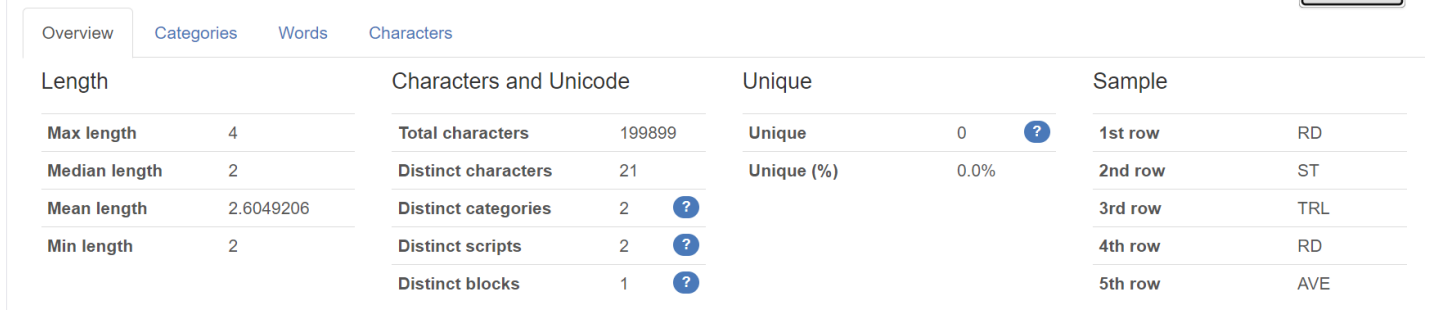
Value	Count	Frequency (%)
preston	2371	2.5%
northwest	2321	2.4%
greenville	2046	2.1%
forest	2016	2.1%
central	1929	2.0%
buckner	1680	1.7%
hill	1496	1.6%
walnut	1435	1.5%
hines	1356	1.4%
harry	1356	1.4%
Other values (859)	78094	81.3%

7. STREET DIRECTION : This column has high correlation. It has 4 distinct values : N,W,S,E stating the 4 directions. It has many missing values 52514 rows. It has same value for min, max and median size : 1.

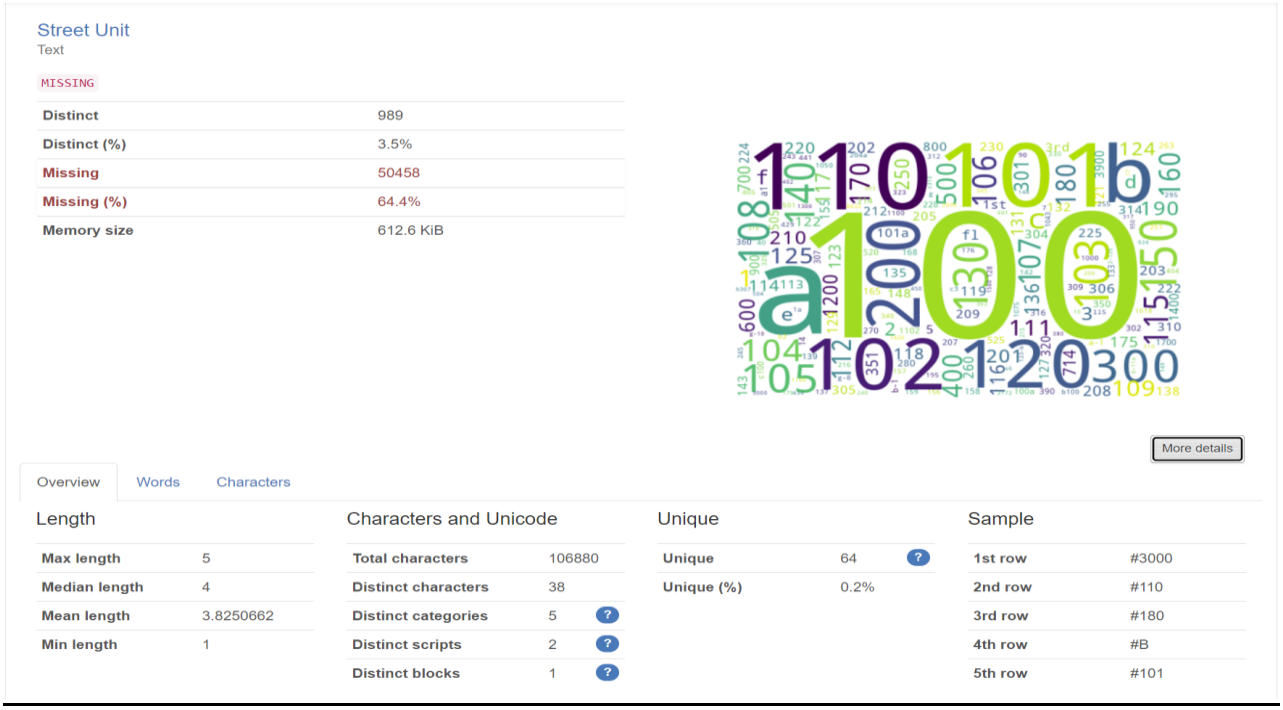


8. STREET TYPE : This column has 19 distinct values i.e. 0.1% of the total. It has 1661 missing values (2.1%) It has max length of 4 and min length of 2 and median length of 2.

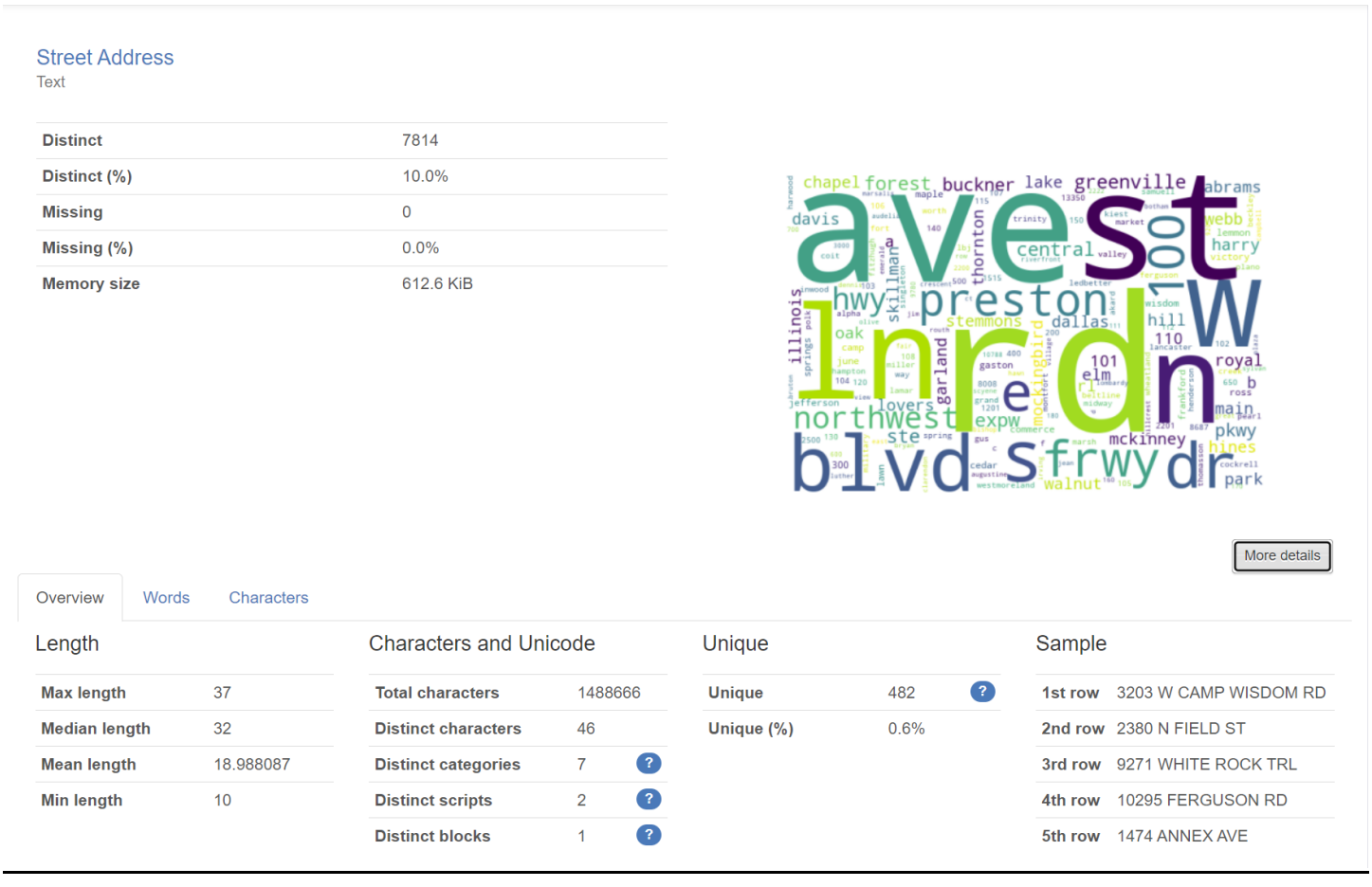




9.STREET UNIT : This column has 989 distinct values I.e. 3.5% of the total values. It has 64.4% values missing, i.e. 50458 values. The length of the column varies from 5 being the max, 1 being the min length and 4 being the median length.



10.STREET ADDRESS : This column represents the street address of each restaurant. It has 7814 distinct values, i.e. 10% of the total number of rows and has no missing values. The length of the street address varies from min. length of 10, and max. length of 37 with median of 32 characters.



11.ZIP CODE : This column has 156 distinct values I.e. 0.2% of the total. There as no missing values present in the column, with min length of 5, max. length of 10 and median length of 5.

Zip Code
Text

Distinct	156
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	612.6 KiB

More details

Overview | Words | Characters

Length

Max length	10
Median length	5
Mean length	5.0152296
Min length	5

Characters and Unicode

Total characters	393194
Distinct characters	11
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

Unique

Unique	12
Unique (%)	< 0.1%

Sample

1st row	75237
2nd row	75021
3rd row	75238
4th row	75228
5th row	75204

Here is a distribution of frequently used zip codes :

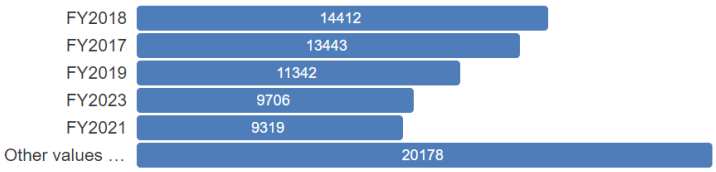
Overview | Words | Characters

Value	Count	Frequency (%)
75201	4678	6.0%
75220	3764	4.8%
75206	3342	4.3%
75243	3117	4.0%
75211	3059	3.9%
75217	3029	3.9%
75231	2968	3.8%
75229	2612	3.3%
75208	2564	3.3%
75228	2541	3.2%
Other values (146)	46726	59.6%

12. **INSPECTION YEAR** : This column has 8 distinct values that constitutes of 0.1% of the total rows. The column same length variations : min, max and median = 6

Inspection Year
Categorical

Distinct	8
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	612.6 KiB



More details

Overview | Categories | Words | Characters

Length

Max length	6
Median length	6
Mean length	6
Min length	6

Characters and Unicode

Total characters	470400
Distinct characters	10
Distinct categories	2
Distinct scripts	2
Distinct blocks	1

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	FY2020
2nd row	FY2020
3rd row	FY2020
4th row	FY2020
5th row	FY2017

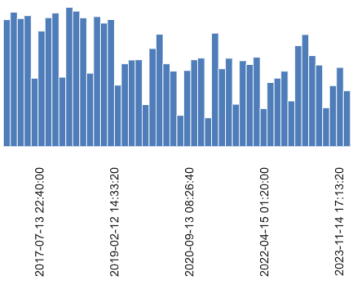
13. **INSPECTION MONTH :** This column has 89 (0.1%) distinct values and no missing values.

Inspection Month

Inspection Month
Date

Distinct	89
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	612.6 KiB

Minimum	2016-10-01 00:00:00
Maximum	2024-02-01 00:00:00



More details

Histogram

14. LATITUDE-LONGITUDE LOCATION : This column represents will be used to show the geographical location of each restaurant. This column has 19299 distinct values, that constitutes of 24.6% of the total rows. There are no missing values associated with this column.

Lat Long Location
Text

Distinct	19299
Distinct (%)	24.6%
Missing	0
Missing (%)	0.0%
Memory size	612.6 KiB

More details

Overview | Words | Characters

Length

Max length	73
Median length	66
Mean length	40.280753
Min length	10

Characters and Unicode

Total characters	3158011
Distinct characters	47
Distinct categories	8
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	9528
Unique (%)	12.2%

Sample

1st row	3203 W CAMP WISDOM RD (32.662584, -96.873446)
2nd row	2380 N FIELD ST (32.78904, -96.806882)
3rd row	9271 WHITE ROCK TRL (32.872855, -96.728807)
4th row	10295 FERGUSON RD (32.83427, -96.673672)
5th row	1474 ANNEX AVE (32.80208, -96.7769)

15. Violation details/Description/Memo : There are total of 75 columns associated with violations details, description, and memo. There are a total of 25 violations and hence 25 sets of violation details, violation description, and violation memo present associated with each restaurant. There are many missing values associated with these columns.