Las Vegas Restaurant Inspections Analysis

Srikar Prayaga

Nationwide®

Data Cleaning and Preparation

Identify the number of null(NA) from every column

RESTAURANT_SERIAL_NUMBER	0
RESTAURANT_PERMIT_NUMBER	0
RESTAURANT_NAME	65
RESTAURANT_LOCATION	200
RESTAURANT_CATEGORY	130
ADDRESS	70
CITY	236
STATE	209
ZIP	59
CURRENT_DEMERITS	216
CURRENT_GRADE	308
EMPLOYEE_COUNT	93
MEDIAN_EMPLOYEE_AGE	34
MEDIAN_EMPLOYEE_TENURE	297
INSPECTION_TIME	183
INSPECTION_TYPE	221
INSPECTION_DEMERITS	254
VIOLATIONS_RAW	165
RECORD_UPDATED	119
LAT_LONG_RAW	15
FIRST_VIOLATION	212
SECOND_VIOLATION	85
THIRD_VIOLATION	61
FIRST_VIOLATION_TYPE	146
SECOND_VIOLATION_TYPE	267
THIRD_VIOLATION_TYPE	173
NUMBER_OF_VIOLATIONS	169
NEXT_INSPECTION_GRADE_C_OR_BELOW	40

Nationwide®

Get a percentage of total number of values which are null. Appears to be <1% on average

```
RESTAURANT_SERIAL_NUMBER          0.000000
RESTAURANT_PERMIT_NUMBER          0.000000
RESTAURANT_NAME                   0.004147
RESTAURANT_LOCATION               0.012761
RESTAURANT_CATEGORY               0.008295
ADDRESS                           0.004466
CITY                              0.015058
STATE                             0.013335
ZIP                               0.003764
CURRENT_DEMERITS                  0.013782
CURRENT_GRADE                     0.019652
EMPLOYEE_COUNT                    0.005934
MEDIAN_EMPLOYEE_AGE               0.002169
MEDIAN_EMPLOYEE_TENURE            0.018950
INSPECTION_TIME                   0.011676
INSPECTION_TYPE                   0.014101
INSPECTION_DEMERITS               0.016206
VIOLATIONS_RAW                    0.010528
RECORD_UPDATED                    0.007593
LAT_LONG_RAW                      0.000957
FIRST_VIOLATION                   0.013526
SECOND_VIOLATION                  0.005423
THIRD_VIOLATION                   0.003892
FIRST_VIOLATION_TYPE              0.009315
SECOND_VIOLATION_TYPE             0.017036

show more (open the raw output data in a text ed

THIRD_VIOLATION_TYPE              0.011038
NUMBER_OF_VIOLATIONS              0.010783
NEXT_INSPECTION_GRADE_C_OR_BELOW  0.002552
dtype: float64
0.9176381584344322
```

Nationwide®

Due to the null values being a very small percentage (<1%) it is feasible to remove them without affecting the outputs

RESTAURANT_SERIAL_NUMBER	0.0
RESTAURANT_PERMIT_NUMBER	0.0
RESTAURANT_NAME	0.0
RESTAURANT_LOCATION	0.0
RESTAURANT_CATEGORY	0.0
ADDRESS	0.0
CITY	0.0
STATE	0.0
ZIP	0.0
CURRENT_DEMERITS	0.0
CURRENT_GRADE	0.0
EMPLOYEE_COUNT	0.0
MEDIAN_EMPLOYEE_AGE	0.0
MEDIAN_EMPLOYEE_TENURE	0.0
INSPECTION_TIME	0.0
INSPECTION_TYPE	0.0
INSPECTION_DEMERITS	0.0
VIOLATIONS_RAW	0.0
RECORD_UPDATED	0.0
LAT_LONG_RAW	0.0
FIRST_VIOLATION	0.0
SECOND_VIOLATION	0.0
THIRD_VIOLATION	0.0
FIRST_VIOLATION_TYPE	0.0
SECOND_VIOLATION_TYPE	0.0

show more (open the raw output data in a text ed

SECOND_VIOLATION_TYPE	0.0
THIRD_VIOLATION_TYPE	0.0
NUMBER_OF_VIOLATIONS	0.0

Nationwide®

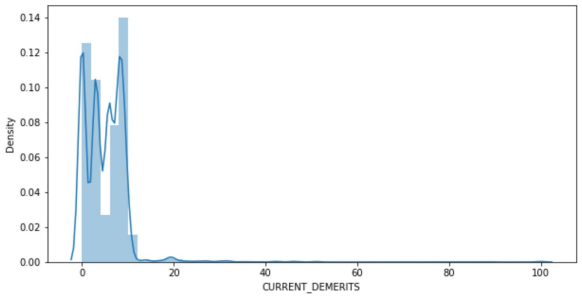
Observe the datatypes of every column to identify if they are appropriate. If not they are casted to the correct type after necessary cleaning!

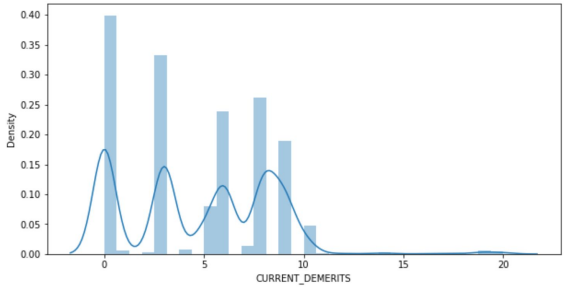
```
RESTAURANT_SERIAL_NUMBER              object
RESTAURANT_PERMIT_NUMBER              object
RESTAURANT_NAME                       object
RESTAURANT_LOCATION                   object
RESTAURANT_CATEGORY                   object
ADDRESS                               object
CITY                                  object
STATE                                 object
ZIP                                   object
CURRENT_DEMERITS                     float64
CURRENT_GRADE                         object
EMPLOYEE_COUNT                       float64
MEDIAN_EMPLOYEE_AGE                  float64
MEDIAN_EMPLOYEE_TENURE               float64
INSPECTION_TIME                       object
INSPECTION_TYPE                       object
INSPECTION_DEMERITS                   object
VIOLATIONS_RAW                        object
RECORD_UPDATED                        object
LAT_LONG_RAW                          object
FIRST_VIOLATION                      float64
SECOND_VIOLATION                     float64
THIRD_VIOLATION                      float64
FIRST_VIOLATION_TYPE                  object
SECOND_VIOLATION_TYPE                 object
THIRD_VIOLATION_TYPE                  object
NUMBER_OF_VIOLATIONS                  object
NEXT_INSPECTION_GRADE_C_OR_BELOW      object
dtype: object
```

Nationwide®

- Use tools like REGEX to remove text which doesn't fit a certain pattern
- Use techniques like 3 Sigma Rule to remove outliers from numerical columns

DA0830561	1
DA0832703	1
DA0574725	1
DAPLVGW0X	1
DA0582330	1
DA1297236	1
DA0983316	1
DA1029996	1
DA0880350	1
DA1135402	1
DA0011120	1
DA0965769	1

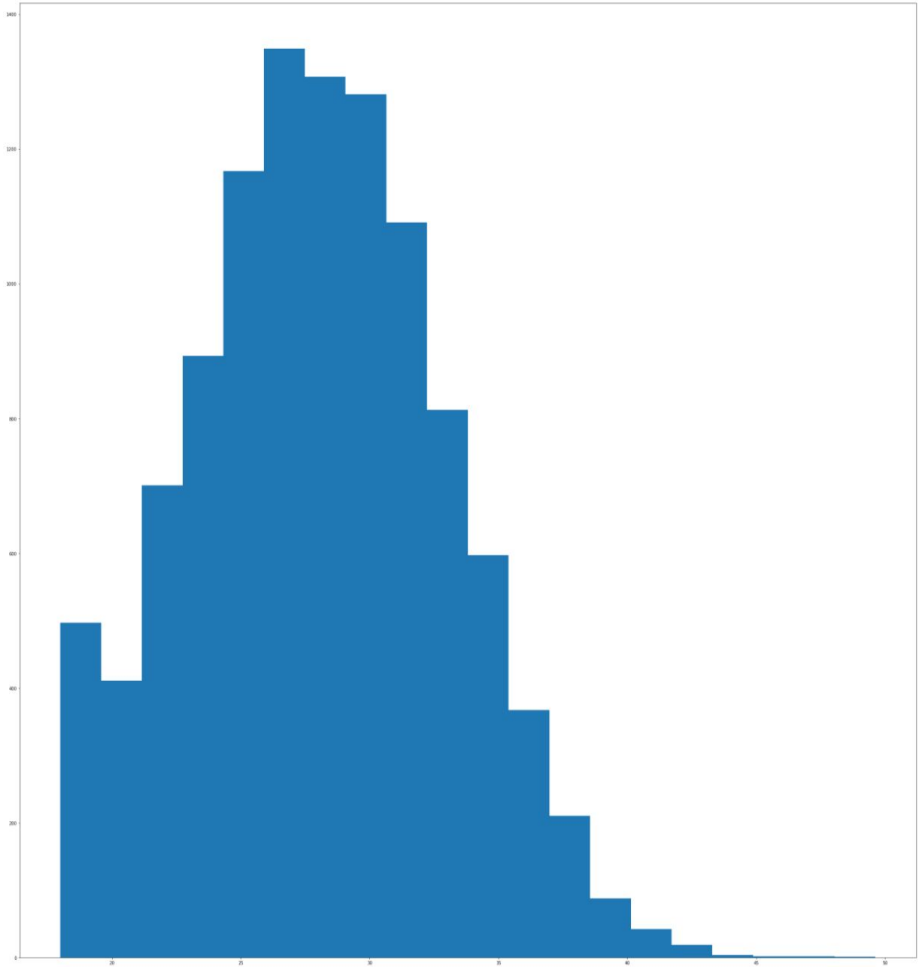




Nationwide®

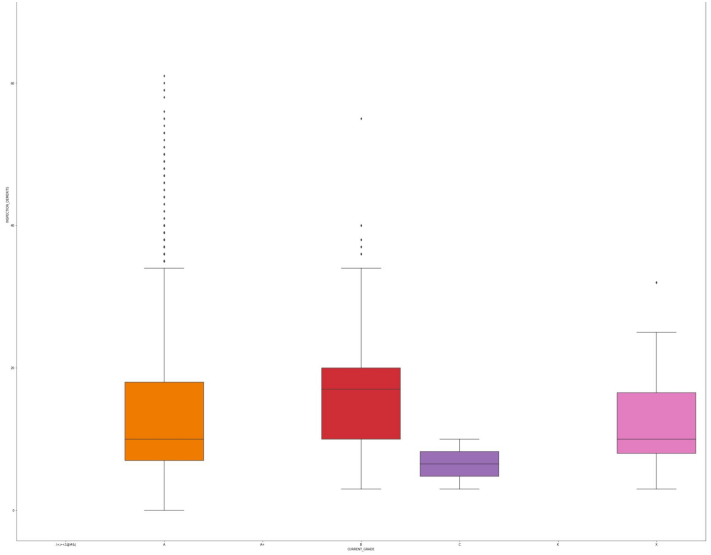
Data Exploration & Transformation

Histogram of MEDIAN_EMPLOYEE_AGE to remove outliers and understand age with respect to other columns in other analysis. Figure shows that most employees are around 30 with a bulk of the staff being less than 30.



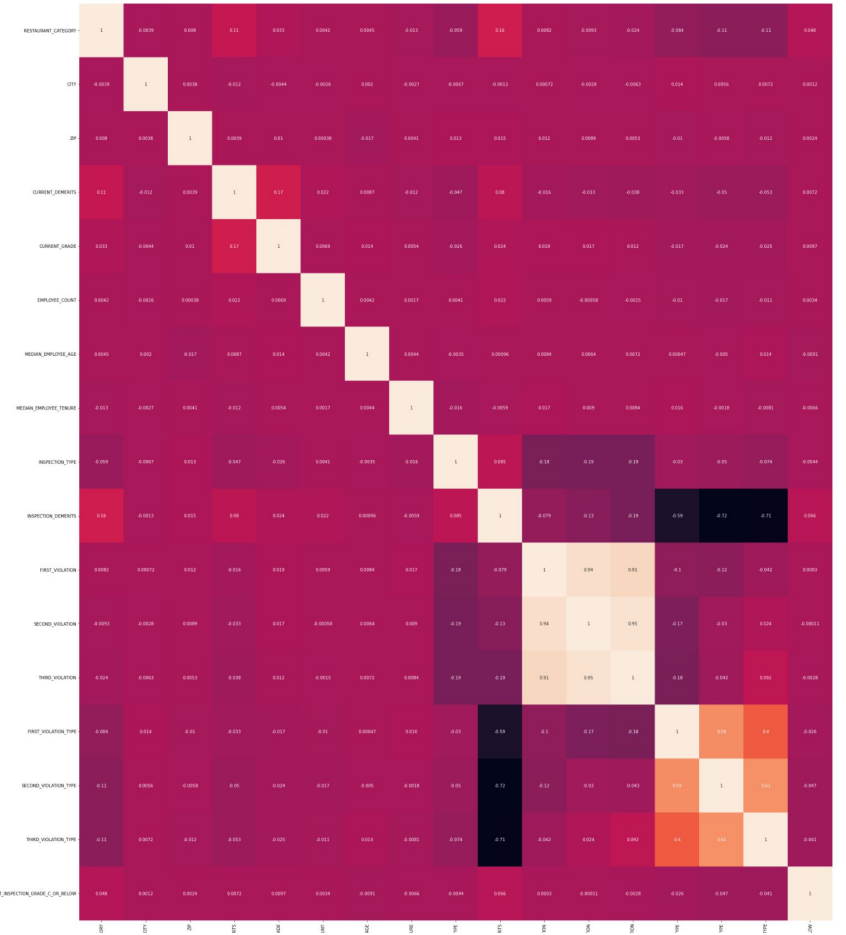
Nationwide®

Box Plot of Demerits vs Current Grade which gives insight into the inverse relationship between both classes.



Nationwide®

- Categorical Ordinal data was Ordinal Encoded while Categorical Nominal data was Label Encoded.
- The coefficient matrix visualizes the relationships between the feature and target class.
- RFE and CHI Squared Test to identify relevant features



Nationwide®

Model Building and Results

- Using Feature Columns:
- 'RESTAURANT_CATEGORY','INSPECTION_DEMERITS','CURRENT_GRADE','NUMBER_OF_VIOLATIONS','EMPLOYEE_COUNT','CITY','ZIP'
- Logistic Regression with an accuracy of 0.85

Accuracy of logistic regression classifier on test set: 0.85

	precision	recall	f1-score	support
0	0.85	1.00	0.92	1840
1	1.00	0.00	0.00	329
accuracy			0.85	2169
macro avg	0.92	0.50	0.46	2169
weighted avg	0.87	0.85	0.78	2169

Nationwide®

- Using Feature Columns:
- 'RESTAURANT_CATEGORY','INSPECTION_DEMERITS','CURRENT_GRADE','NUMBER_OF_VIOLATIONS','EMPLOYEE_COUNT','CITY','ZIP'
- Random Forest Classifier with an accuracy rate of 0.83

Accuracy of random forest classifier on test set: 0.83

	precision	recall	f1-score	support
0	0.85	0.97	0.91	1840
1	0.24	0.05	0.08	329
accuracy			0.83	2169
macro avg	0.55	0.51	0.50	2169
weighted avg	0.76	0.83	0.78	2169

Nationwide®

- Using Feature Columns:
- 'RESTAURANT_CATEGORY','INSPECTION_DEMERITS','CURRENT_GRADE','NUMBER_OF_VIOLATIONS','EMPLOYEE_COUNT','CITY','ZIP'
- Naive Bayes Classifier with an accuracy rate of 0.82

Accuracy of Naive Bayes classifier on test set: 0.82

	precision	recall	f1-score	support
0	0.86	0.95	0.90	1840
1	0.26	0.10	0.15	329
accuracy			0.82	2169
macro avg	0.56	0.53	0.52	2169
weighted avg	0.76	0.82	0.78	2169

Nationwide®

- Using Feature Columns:
- 'RESTAURANT_CATEGORY','INSPECTION_DEMERITS','CURRENT_GRADE','NUMBER_OF_VIOLATIONS','EMPLOYEE_COUNT','CITY','ZIP'
- Decision Tree Classifier with an accuracy rate of 0.73

Accuracy of Decision Tree classifier on test set: 0.73

	precision	recall	f1-score	support
0	0.85	0.82	0.84	1840
1	0.17	0.21	0.19	329
accuracy			0.73	2169
macro avg	0.51	0.52	0.51	2169
weighted avg	0.75	0.73	0.74	2169

- Using Feature Columns:
- 'RESTAURANT_CATEGORY','INSPECTION_DEMERITS','CURRENT_GRADE','NUMBER_OF_VIOLATIONS','EMPLOYEE_COUNT','CITY','ZIP'
- Gradient Boosting Classifier with an accuracy rate of 0.85

Accuracy of Gradient Boosting classifier on test set: 0.85

	precision	recall	f1-score	support
0	0.85	1.00	0.92	1840
1	0.00	0.00	0.00	329
accuracy			0.85	2169
macro avg	0.42	0.50	0.46	2169
weighted avg	0.72	0.85	0.78	2169

- Using Feature Columns:
- 'RESTAURANT_CATEGORY','INSPECTION_DEMERITS','CURRENT_GRADE','NUMBER_OF_VIOLATIONS','EMPLOYEE_COUNT','CITY','ZIP'
- KNN Classifier with an accuracy rate of 0.83

Accuracy of knn classifier on test set: 0.83

	precision	recall	f1-score	support
0	0.85	0.97	0.90	1840
1	0.20	0.05	0.07	329
accuracy			0.83	2169
macro avg	0.52	0.51	0.49	2169
weighted avg	0.75	0.83	0.78	2169

Final Model: Gaussian Naive Bayes with Accuracy rate of 0.82

```
Accuracy of MultinomialNB classifier on test set: 0.74
              precision    recall  f1-score   support

           0       0.85      0.84      0.85      1840
           1       0.17      0.19      0.18       329

    accuracy                           0.74      2169
   macro avg       0.51      0.51      0.51      2169
weighted avg       0.75      0.74      0.74      2169

Accuracy of BernoulliNB classifier on test set: 0.85
              precision    recall  f1-score   support

           0       0.85      1.00      0.92      1840
           1       0.00      0.00      0.00       329

    accuracy                           0.85      2169
   macro avg       0.42      0.50      0.46      2169
weighted avg       0.72      0.85      0.78      2169

Accuracy of GaussianNB classifier on test set: 0.82
              precision    recall  f1-score   support

           0       0.86      0.95      0.90      1840
           1       0.26      0.10      0.15       329

show more (open the raw output data in a text editor) ...

    accuracy                           0.82      2169
   macro avg       0.56      0.53      0.52      2169
weighted avg       0.76      0.82      0.78      2169
```

Questions?