Machine Learning - Part I

Simone Genovese

21 October 2024

Contents

1 Introduction **2**
- 1.1 NaN exploration . . . 2
- 1.2 Value exploration . . . 3
  - 1.2.1 Zip Code . . . 3
  - 1.2.2 District Name . . . 4
  - 1.2.3 Industry Code . . . 5
  - 1.2.4 Industry Code Description (ICD) . . . 7
  - 1.2.5 Medical Fee Region (MFR) . . . 8
  - 1.2.6 County of Injury (C of I) . . . 10
- 1.3 Correlation Analysis . . . 11

1 Introduction

13. object County of Injury: Name of the New York County where the injury occurred.

15. object District Name: Name of the WCB district office that oversees claims for that region or area of the state.

19. float64 Industry Code: NAICS code and descriptions.

20. object Industry Code Description: 2-digit NAICS industry code description used to classify businesses according to their economic activity.

21. object Medical Fee Region Approximate region where the injured worker would receive medical service

29. object Zip Code The reported ZIP code of the injured worker's home address

1.1 NaN exploration

Among the geographical data, here are percentages of missing records:

	%Records with at least one NAN	%Records with only NAN
train	9.693650	3.276487
test	4.498282	0.000000

In the training data, almost 10% of records have at least one missing value, and 1/3 of them have no values at all.

In the test data, the proportion is way lower and we can find at least one geographical value in each record.

In detail, for every variable:

	train	test
Zip Code	0.081018	0.049854
District Name	0.032765	0.000000
County of Injury	0.032765	0.000000
Industry Code	0.049544	0.019939
Industry Code Description	0.049544	0.019939
Medical Fee Region	0.032765	0.000000

The following data frame shows the proportions (with respect to the set of incomplete records) of records that present pairwise or triplet-wise all NaN in the rows specified. On the diagonal appear the contribution of NaN of the single variable. It wants to measure a "causal correlation" in presence of NAN among different columns.

	Zip Code	District Name	C of I	Industry Code	ICD	MFR
Zip Code	0.84	0.34	0.34	0.35	0.35	0.34
District Name	NaN	0.34	0.34	0.34	0.34	0.34
C of I	NaN	Zip Code 0.34	0.34	0.34	0.34	0.34
Industry Code	NaN	Zip Code 0.34	District Name 0.34	0.51	0.51	0.34
ICD	NaN	Zip Code 0.34	District Name 0.34	C of I 0.34	0.51	0.34
MFR	NaN	Zip Code 0.34	District Name 0.34	C of I 0.34	Industry Code 0.34	0.34

As far as we can see, there seems not to be such a relationship, but a note for one pair: half of the non registered Industry Codes doesn't have the Industry Code Description registered as well. As a further proof for the previous analysis, about one third of the records are missing (0.34 spammed everywhere).

1.2 Value exploration

In the following section we proceed with the exploration of the actual values inside variables, one per one. We'll figure out:

1. summary of values (.describe())
2. percentage of missing values
3. percentage of numeric values
4. whether the values are recognized as 'object' because of presence of NaNs, but it's actually numeric
5. distributions
6. "outliers" (we'll see that they cannot be considered thus. Boxplots are not relevant for these kind of variables

1.2.1 Zip Code

The Zip Code variable is the most tricky one because of the amount of unique values and also their format.

	train	test
count	545389	368633
unique	8286	6276
top	11236	11368
freq	3398	2068
is_actually_numeric	False	False
numeric mode	('11236', 3398)	('11368', 2068)
non numeric mode	('V6T1Z', 41)	('M3K2C', 45)

- Not all the Zip Codes are fully numeric. There are also alphanumeric values even without a defined schema (namely, 2 letters and 3 digits). From a web research, it seems that there cannot exist postal codes with letters, or at least there's no correspondence with the 5-digits format ones.
- The percentage of non-numeric zip codes is under the 1% of the total. NaN reach the 8.1%, all the other values are numeric as the USA zip code wants by default.
- in training data 3567 zip codes appear only once, in test data the number is 2783. It can be defined a threshold of number of appearance to consider whether a value is an outlier or not*

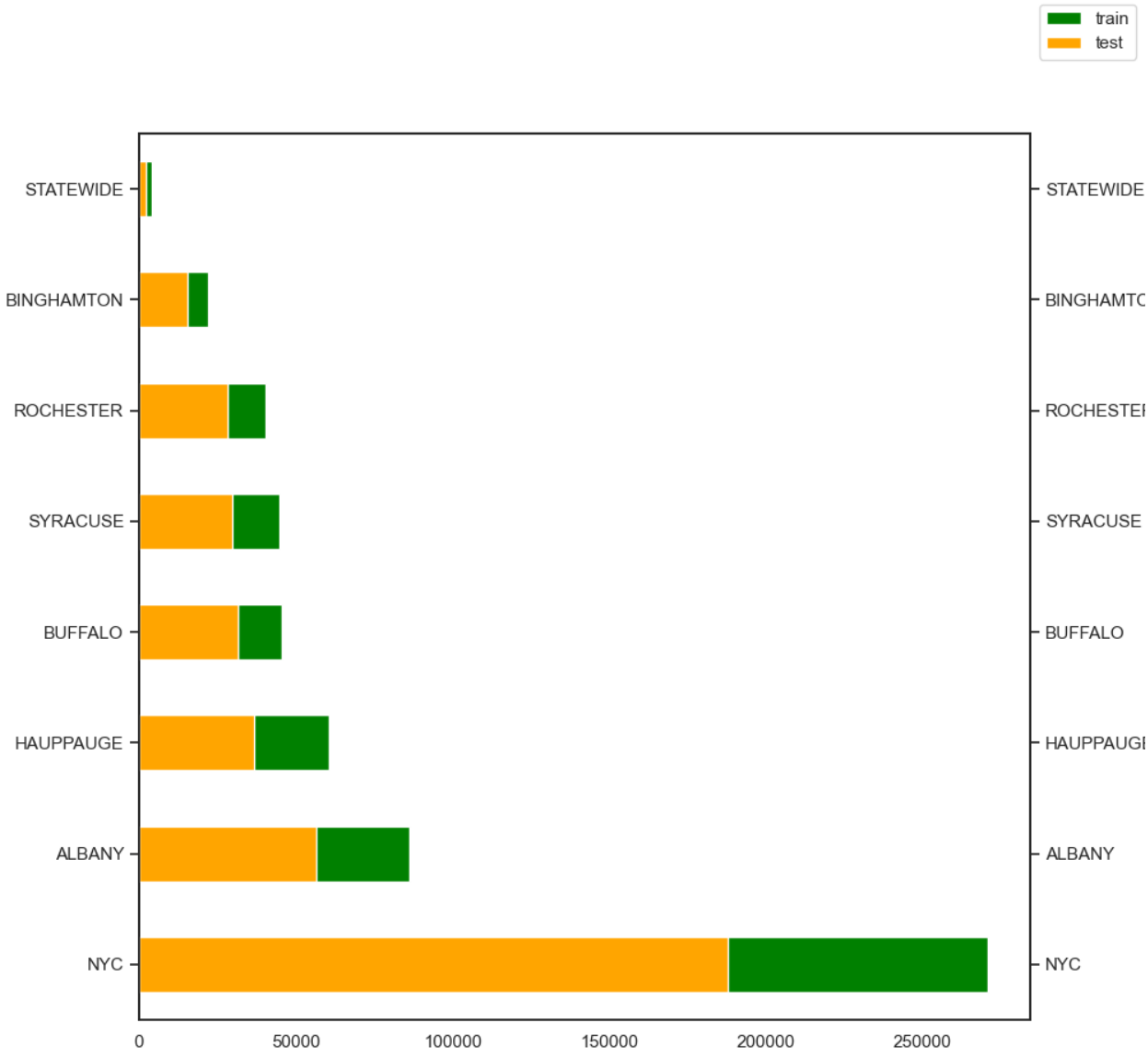
*the following output exposes a trial.
Threshold of 1 : 4719 different zip codes
Threshold of 2 : 3598 different zip codes
Threshold of 3 : 3037 different zip codes
Threshold of 4 : 2720 different zip codes
Threshold of 5 : 2520 different zip codes
Threshold of 6 : 2363 different zip codes
Threshold of 7 : 2254 different zip codes
Threshold of 8 : 2161 different zip codes
Threshold of 9 : 2087 different zip codes
Threshold of 10 : 2026 different zip codes
Threshold of 11 : 1962 different zip codes
Threshold of 12 : 1903 different zip codes
Threshold of 13 : 1857 different zip codes
Threshold of 14 : 1818 different zip codes

1.2.2 District Name

	train	test
count	574026	387975
unique	8	8
top	NYC	NYC
freq	270779	187972
is_actually_numeric	False	False
non numeric mode	('NYC', 270779)	('NYC', 187972)

District Name	STATEWIDE
train outliers	3976
test outliers	2374

- District names are actual names in string format
- only 8 types of districts are recorded, recognizing the variable as 'category'. No numeric value appears
- NYC is the most rated district with a relevant discrepancy in proportion with respect to the other ones
- there's no reason to recognize the less frequent value as an outlier

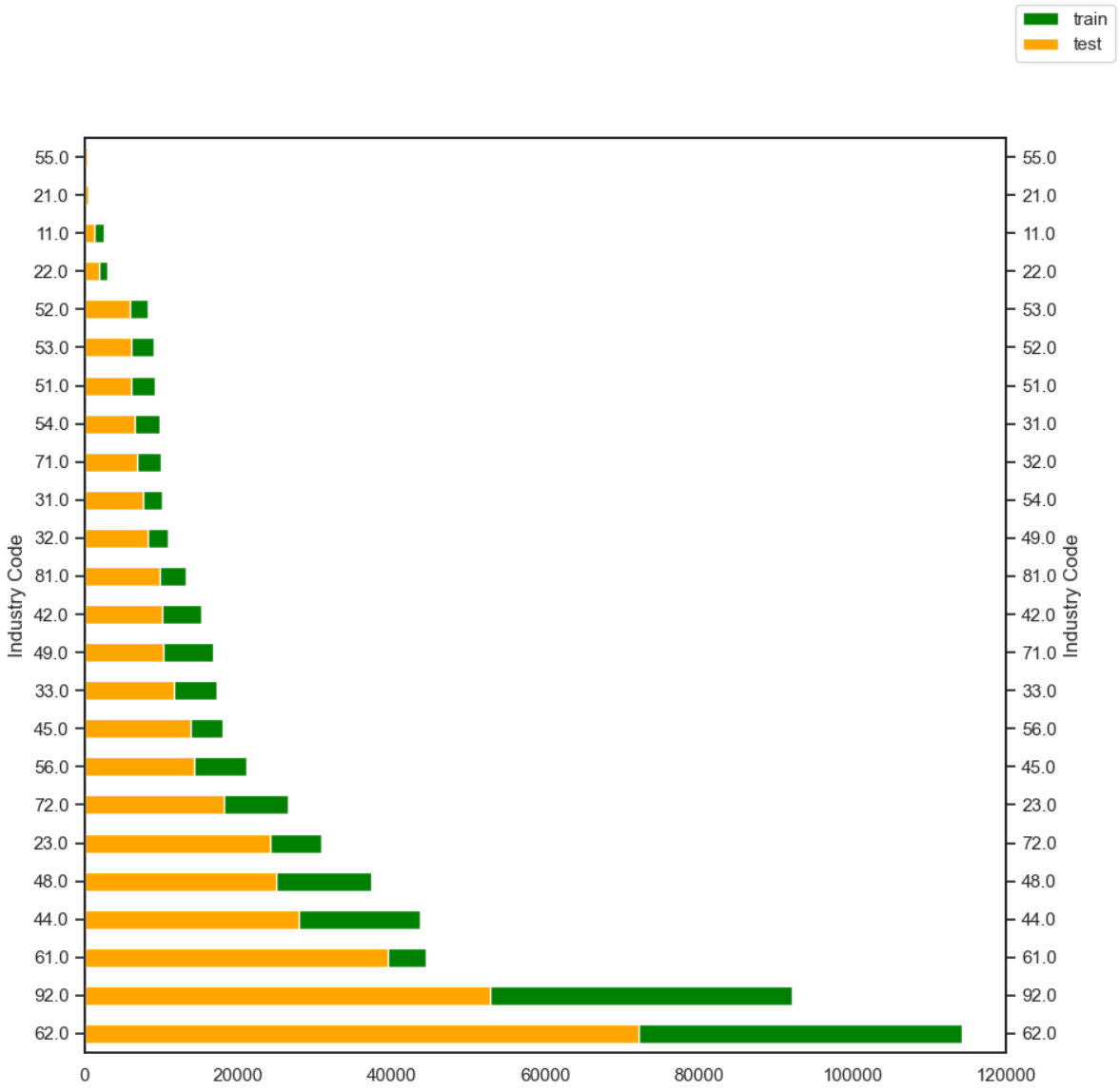


1.2.3 Industry Code

	Industry Code	Industry Code
count	564068	380239
unique	24	24
top	62	62
freq	114339	72207
%_na	4.95	1.99
non numeric mode	(62.0, 114339)	(62.0, 72207)

Industry Code	55
train outliers	370
test outliers	294

- Industrial codes are recognized as float but they clearly represent discrete and categorical values
- only 8 types of districts are recorded, recognizing the variable as 'category'. No numeric value appears
- 62 and 92 are the most frequent values proportionally dominant
- there's no reason to recognize the less frequent value as an outlier

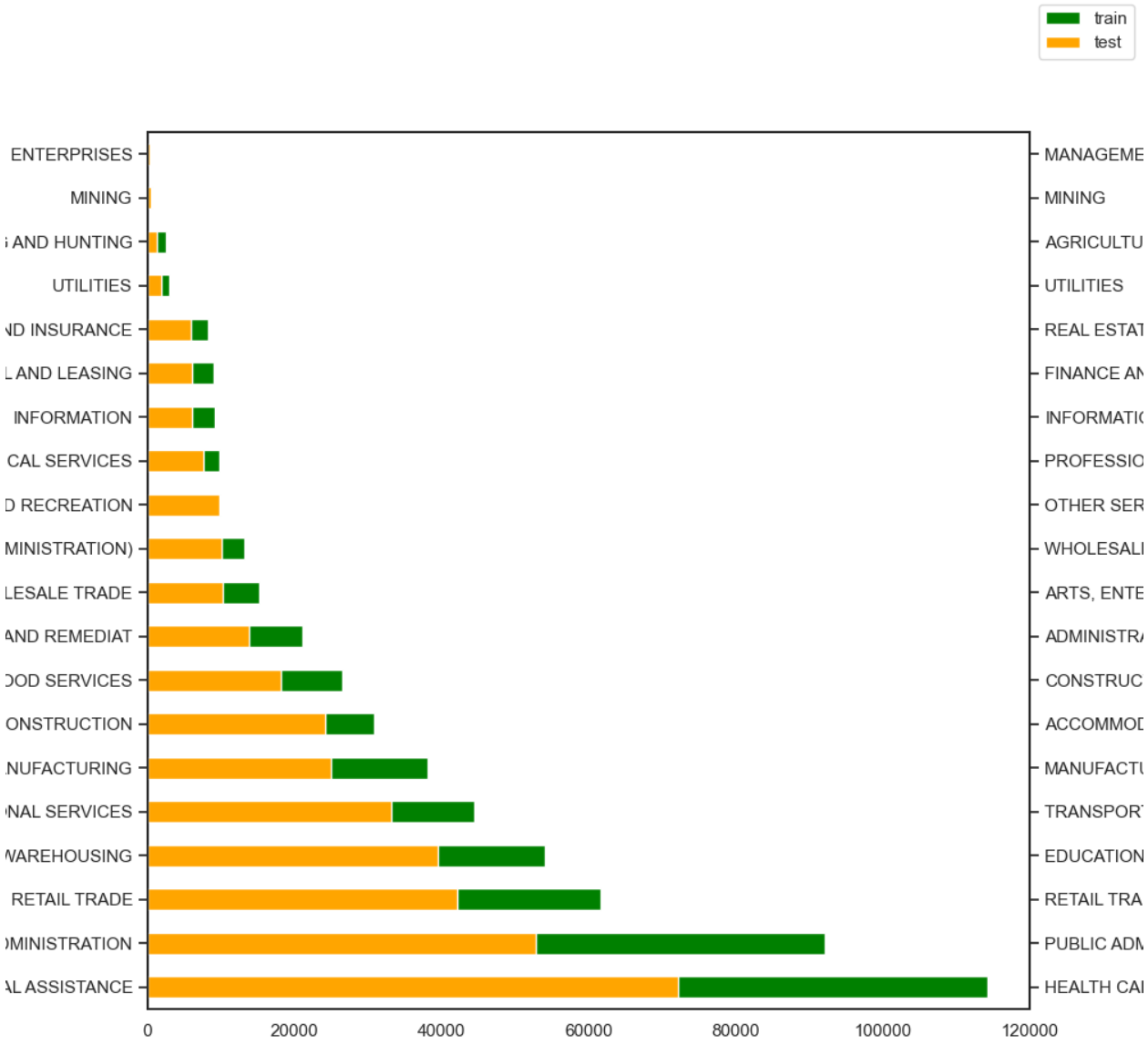


1.2.4 Industry Code Description (ICD)

	ICD	ICD
count	564068	380239
unique	20	20
top	HEALTH CARE AND SOCIAL ASSISTANCE	HEALTH CARE AND SOCIAL ASSIS
freq	114339	72207
is_actually_numeric	False	False
non numeric mode	('HEALTH CARE AND SOCIAL ASSISTANCE', 114339)	('HEALTH CARE AND SOCIAL ASSI

ICD	MANAGEMENT OF COMPANIES AND ENTERPRISES
train outliers	370
test outliers	294

- descriptions are standardized strings (strings from a ready-made list of categories)
- only 20 types of districts are recorded, recognizing the variable as 'category'. No numeric value appears
- there are no actual dominant values, the distribution looks to follow an exponential behavior
- there's no reason to recognize the less frequent value as an outlier

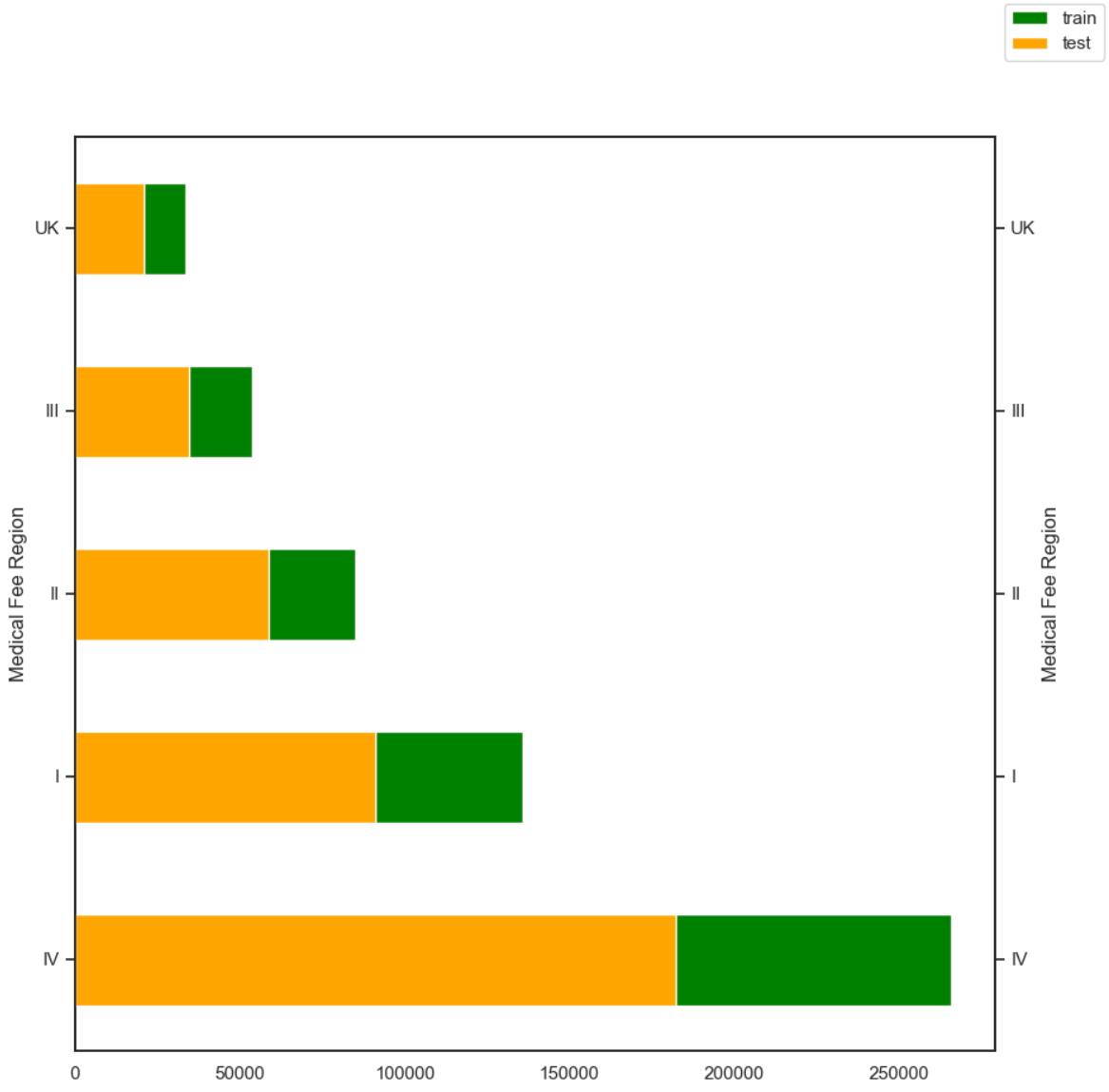


1.2.5 Medical Fee Region (MFR)

	MFR	MFR
count	574026	387975
unique	5	5
top	IV	IV
freq	265981	182276
$is_actually_numeric$	False	False
non numeric mode	('IV', 265981)	('IV', 182276)

MFR	UK
train outliers	33473
test outliers	20977

- fee regions are determined by level (from I to IV) and UK in case. 5 categories in total
- the IV region looks slightly to dominate
- there's no reason to recognize the less frequent value as an outlier

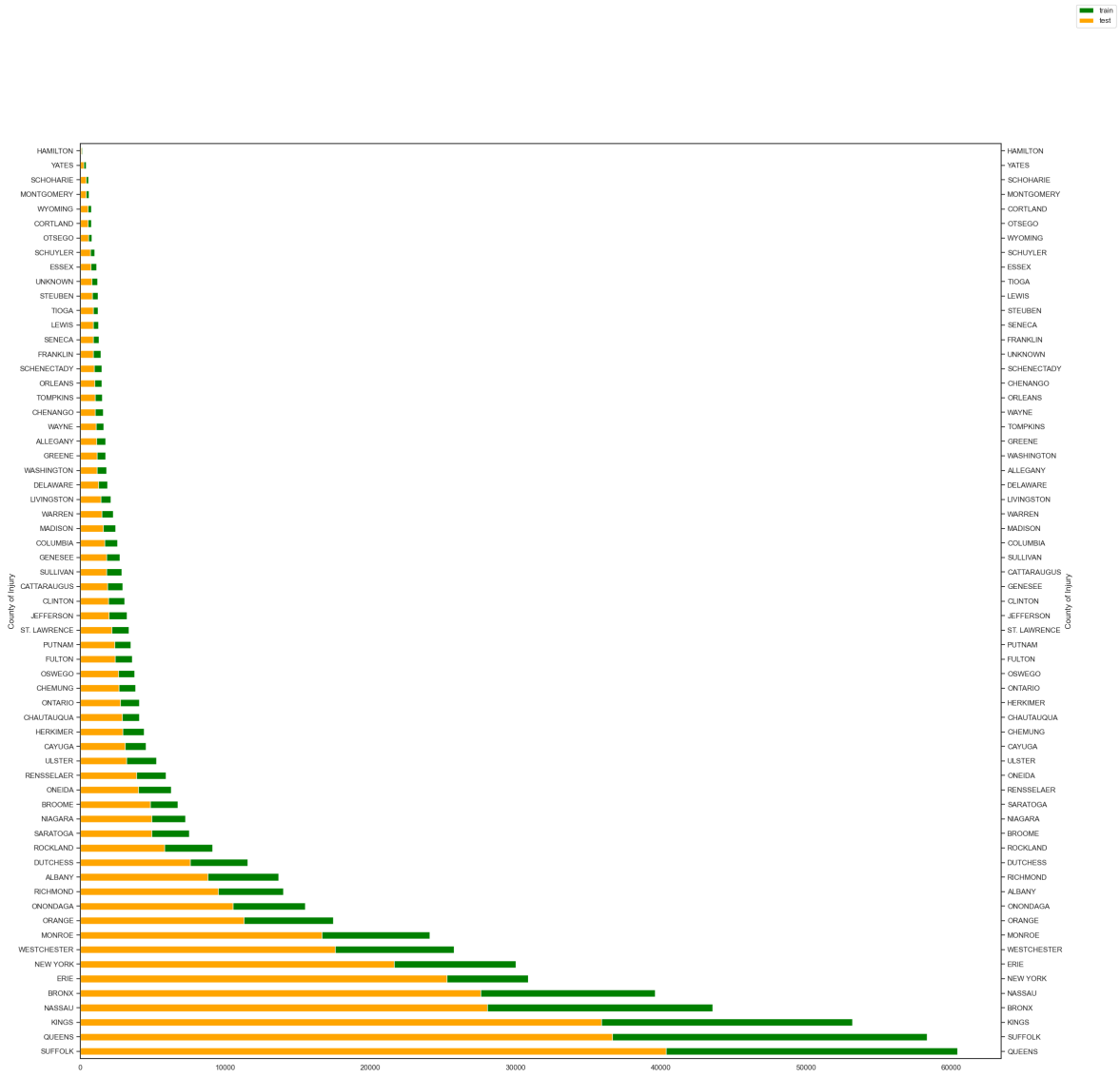


1.2.6 County of Injury (C of I)

	C of I	C of I
count	574026	387975
unique	63	63
top	SUFFOLK	QUEENS
freq	60430	40358
is_actually_numeric	False	False
non numeric mode	('SUFFOLK', 60430)	('QUEENS', 40358)

C of I	HAMILTON
train outliers	134
test outliers	97

- counties are recorded by their full names, 63 occurred
- no numeric neither strange values inside
- frequencies clearly follow an exponential distribution, with a few differences between train and test data (e.g. QUEENS and SUFFOLK are switched)
- there's no reason to recognize the less frequent value as an outlier



1.3 Correlation Analysis

A simple chi square indipendency test doesn't give reliable results, better to go for a CramerV test. There are no regression correlations (like Pearson) suitable for our purposes since data are fully categorical.

train	Zip Code	District Name	C of I	Industry Code	ICD	MFR
Zip Code	1.00	0.90	0.95	0.20	0.21	1.00
District Name	-	1.00	0.91	0.12	0.11	0.49
C of I	-	-	1.00	0.09	0.10	0.72
Industry Code	-	-	-	1.00	1.00	0.14
ICD	-	-	-	-	1.00	0.13
MFR	-	-	-	-	-	1.00

test	Zip Code	District Name	C of I	Industry Code	ICD	MFR
Zip Code	1.00	0.91	0.96	0.22	0.23	1.00
District Name	-	1.00	0.92	0.12	0.11	0.49
C of I	-	-	1.00	0.10	0.10	0.74
Industry Code	-	-	-	1.00	1.00	0.16
ICD	-	-	-	-	1.00	0.15
MFR	-	-	-	-	-	1.00

- A high correlation appears among Zip Code and County of Injury. This may lead to assume that most of people work in the neighborhoods they live in
- A high correlation appears among Zip Code and District name. This may lead to assume that most of people claim for the injury in the District they live in
- A high correlation appears among County of Injury and District name. This may lead to assume that there is a very small number of claim offices in each county
- Industry codes and their description are almost in a one-to-one relationship. "almost" because sometimes more Industry Codes refer to the same description. One variable can be dropped, I suggest to keep the Code in order to better classify the type of job rather than macro-areas. For more information: https://www.naics.com/search-naics-codes-by-industry/
- In the majority of cases, the medical fee region of healing is located in the county where the injury occurred
- both train and test data show a similar behavior

The similarity among the two results is more evident when computing the delta between the two matrices

delta	Zip Code	District Name	C of I	Industry Code	ICD	MFR
Zip Code	0.000	0.009	0.019	0.016	0.018	0.000
District Name	-	0.000	0.007	0.003	0.003	0.004
C of I	-	-	0.000	0.008	0.009	0.017
Industry Code	-	-	-	0.000	0.000	0.016
ICD	-	-	-	-	0.000	0.016
MFR	-	-	-	-	-	0.000