shudhi_describe

April 27, 2018

0.1 How to use: "shudhi_describe"

```
In [3]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        import warnings
        import time
        from scipy.stats import zscore
        from datetime import datetime
        from IPython.display import display, HTML
```

0.2 Import Shudhi

```
In [1]: from shudhi_describe import shudhi_describe
```

0.3 Import Datasets: we have a small and a big dataset

```
In [4]: df_four = pd.read_json('foursquare_test.json', orient='records')

In [5]: df_four.shape

Out[5]: (400, 11)

In [6]: df_complaint = pd.read_csv('Top_5_complaints.csv')

In [7]: df_complaint.shape

Out[7]: (376062, 36)
```

1 Run Shudhi Describe with various params

1.0.1 Function: shudhi_describe(df_train, cols= [None], empty_missing= False, plot=True, target= None)

1.0.2 Params:

1. df_train -> the dataset as a Pandas DataFrame
2. cols -> a list of columns (Default all columns)
3. empty_missing -> if True, empty cell ("") is considered as a np.nan
4. plot -> If False, plots are not shown
5. Target -> If givn a column name(str), bivariate scatter plots will be displayed too

1.1 The Small dataset: takes ~2 s

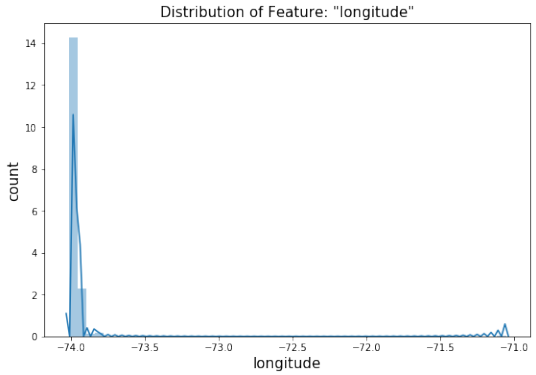
```
In [8]: start_time = time.time()
        shudhi_describe(df_four)
        print("Runtime: %s seconds" % round((time.time() - start_time)))
```

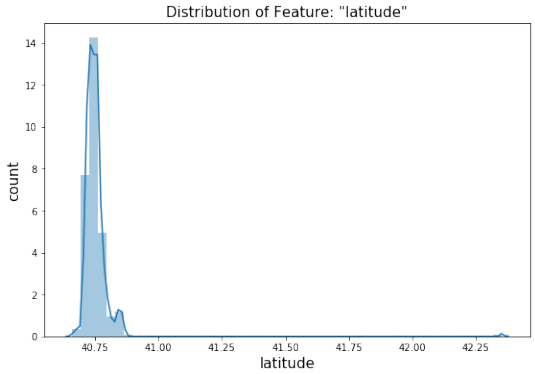
SUMMARY STATISTICS

	Feature	Feature Type	count	# Unique	# Missing	# Outliers	mean	median	min	max
0	country	String/Object	400	1	0					
1	id	String/Object	400	400	0					
2	latitude	Real Value	400	399	0	1	40.75	40.74	40.6605	42.3531
3	locality	String/Object	400	5	0					
4	longitude	Real Value	400	399	0	1	-73.97	-73.99	-74.0159	-71.0541
5	name	String/Object	400	386	0					
6	phone	String/Object	400	199	181					
7	postal_code	Real Value saved as string	400	53	0					
8	region	String/Object	400	1	0					
9	street_address	String/Object	400	310	0					
10	website	Date/Time saved as string	400	69	0					

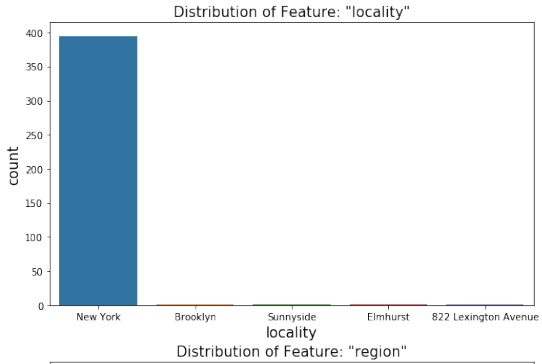
PLOTS

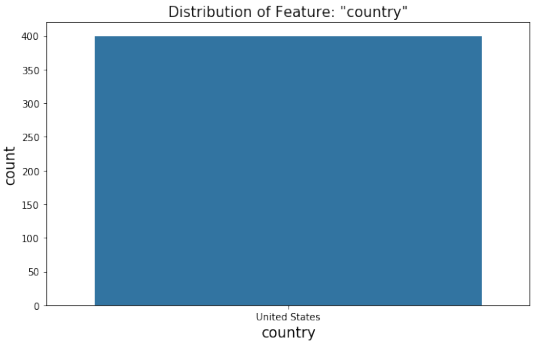
UNIVATIATE PLOTS: Continuous Features

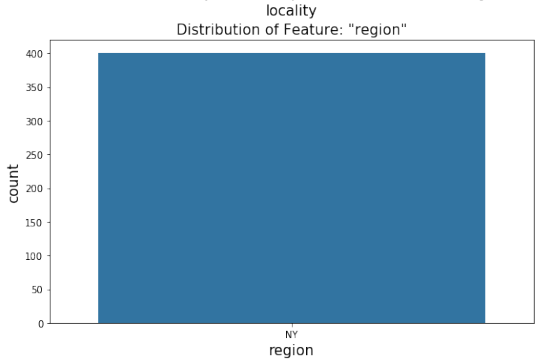


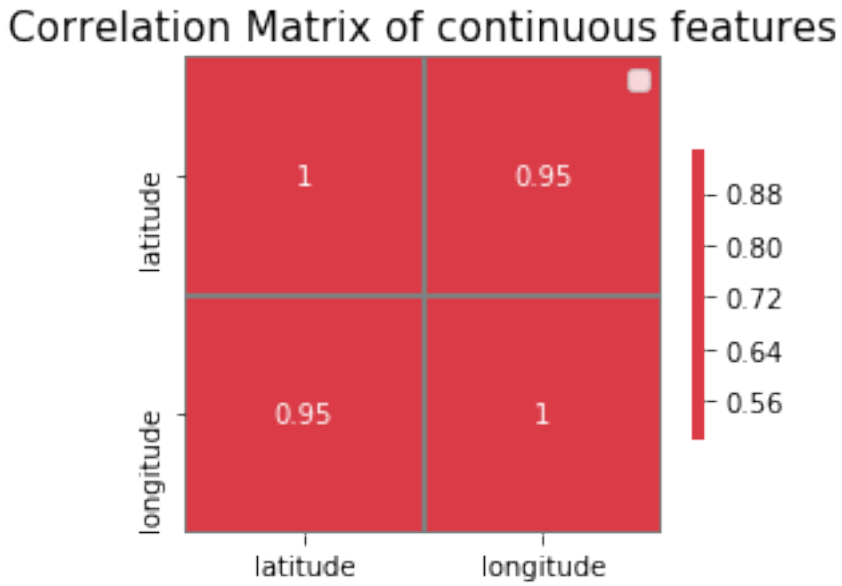


UNIVATIATE PLOTS: Categorical Features










```
Runtime: 2 seconds
```

1.2 The Big Dataset : takes ~ 60 s

```
In [9]: start_time = time.time()
        shudhi_describe(df_complaint)
        print("Runtime: %s seconds" % round((time.time() - start_time)))
```

SUMMARY STATISTICS

	Feature	Feature Type	count	\
0	Unique Key	Integer	376062	
1	Created Date	Date/Time saved as string	376062	
2	Closed Date	Date/Time saved as string	376062	
3	Agency	String/Object	376062	
4	Agency Name	String/Object	376062	
5	Complaint Type	String/Object	376062	
6	Descriptor	String/Object	376062	
7	Location Type	String/Object	376062	
8	Incident Zip	Real Value	376062	
9	Incident Address	String/Object	376062	
10	Street Name	String/Object	376062	
11	Cross Street 1	String/Object	376062	
12	Cross Street 2	String/Object	376062	
13	Intersection Street 1	String/Object	376062	

14	Intersection Street 2	String/Object	376062	
15	Address Type	String/Object	376062	
16	City	String/Object	376062	
17	Landmark	String/Object	376062	
18	Status	String/Object	376062	
19	Due Date	Date/Time saved as string	376062	
20	Resolution Description	String/Object	376062	
21	Resolution Action Updated Date	Date/Time saved as string	376062	
22	Community Board	String/Object	376062	
23	Borough	String/Object	376062	
24	X Coordinate (State Plane)	Real Value	376062	
25	Y Coordinate (State Plane)	Real Value	376062	
26	Park Facility Name	String/Object	376062	
27	Park Borough	String/Object	376062	
28	Vehicle Type	Real Value	376062	
29	Garage Lot Name	Real Value	376062	
30	Ferry Terminal Name	Real Value	376062	
31	Latitude	Real Value	376062	
32	Longitude	Real Value	376062	
33	Location	String/Object	376062	
34	year	Integer	376062	
35	month	Integer	376062	

	# Unique	# Missing	# Outliers	mean	median	min	\
0	376062	0	0	3.34056e+07	3.37504e+07	1.95111e+07	
1	334662	0					
2	273775	2239					
3	7	0					
4	39	0					
5	5	0					
6	62	0					
7	20	31986					
8	209	2370	0	10795.9	11203	0	
9	151287	34479					
10	8597	34479					
11	9418	122944					
12	9406	123443					
13	4951	341669					
14	4860	341916					
15	5	1292					
16	90	2356					
17	42	375970					
18	5	0					
19	217043	139610					
20	98	327					
21	274121	1152					
22	76	0					
23	6	0					

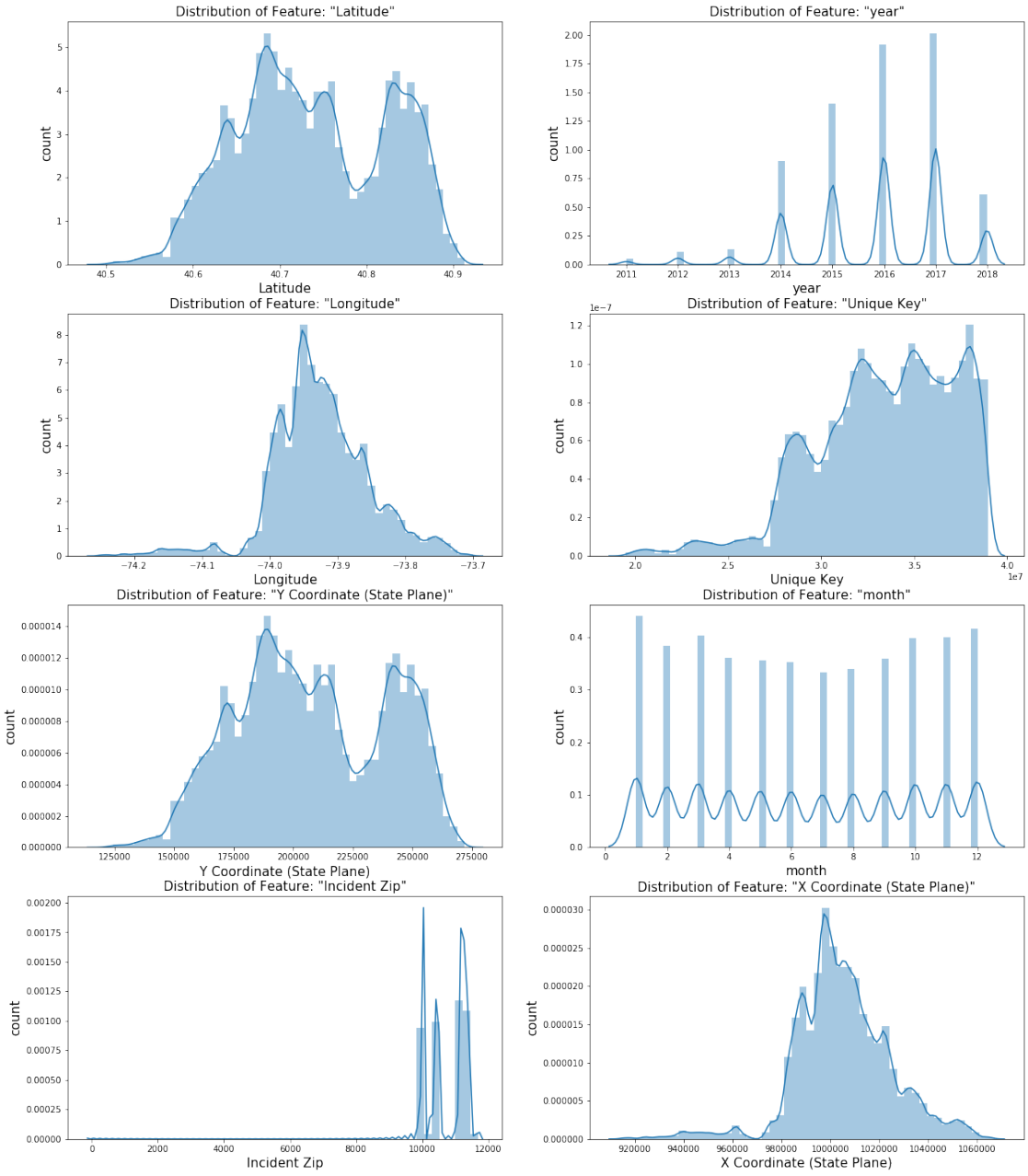
24	69741	2996	0	1.00514e+06	1.0039e+06	913495
25	85139	2996	0	207940	204824	121212
26	1	0				
27	6	0				
28	0	376062	0			
29	0	376062	0			
30	0	376062	0			
31	164759	2996	0	40.74	40.73	40.4991
32	164663	2996	0	-73.92	-73.93	-74.2544
33	165828	2996				
34	8	0	0	2015.85	2016	2011
35	12	0	0	6.47	6	1

	max
0	3.89483e+07
1	
2	
3	
4	
5	
6	
7	
8	11697
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	
24	1.06717e+06
25	271876
26	
27	
28	
29	
30	
31	40.9129
32	-73.7008
33	

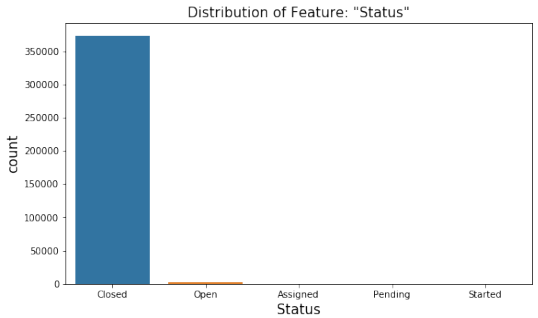
```
34          2018
35            12
```

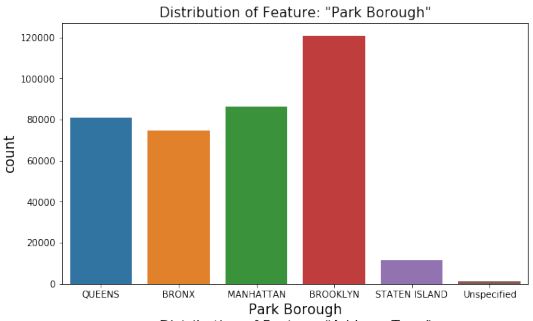
PLOTS

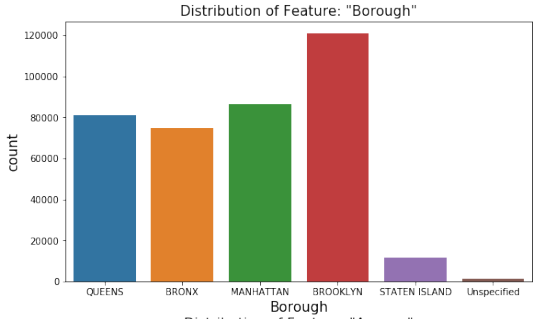
UNIVATIATE PLOTS: Continuous Features

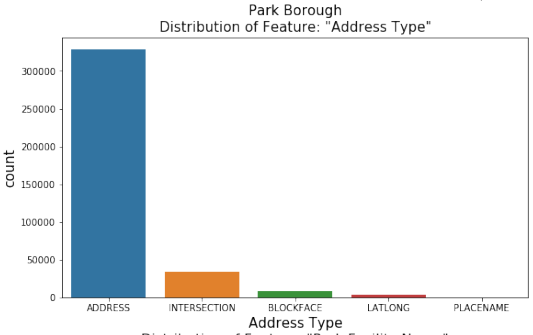


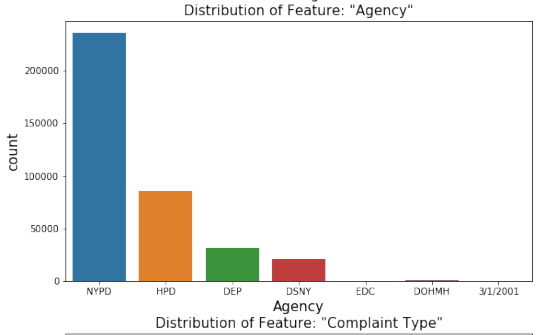
UNIVATIATE PLOTS: Categorical Features

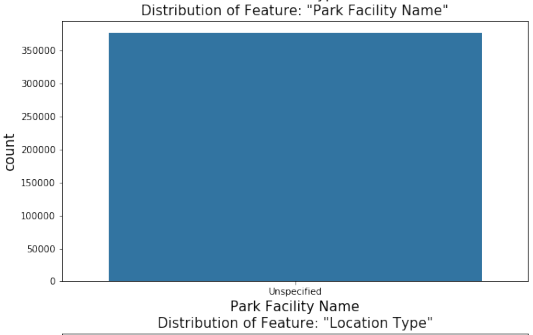


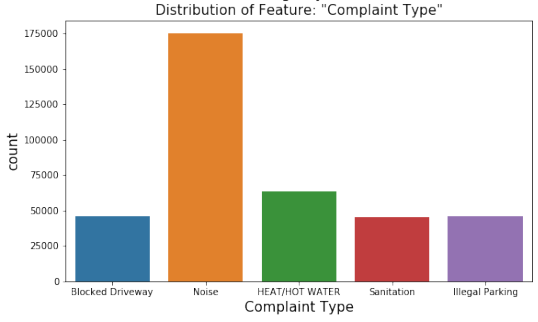


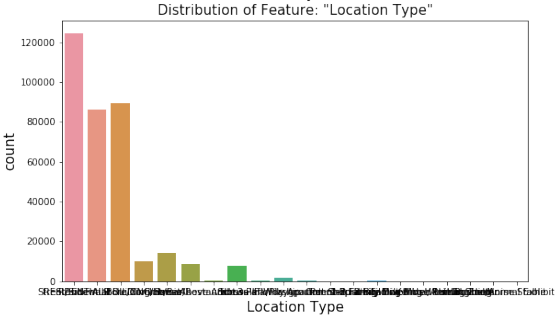


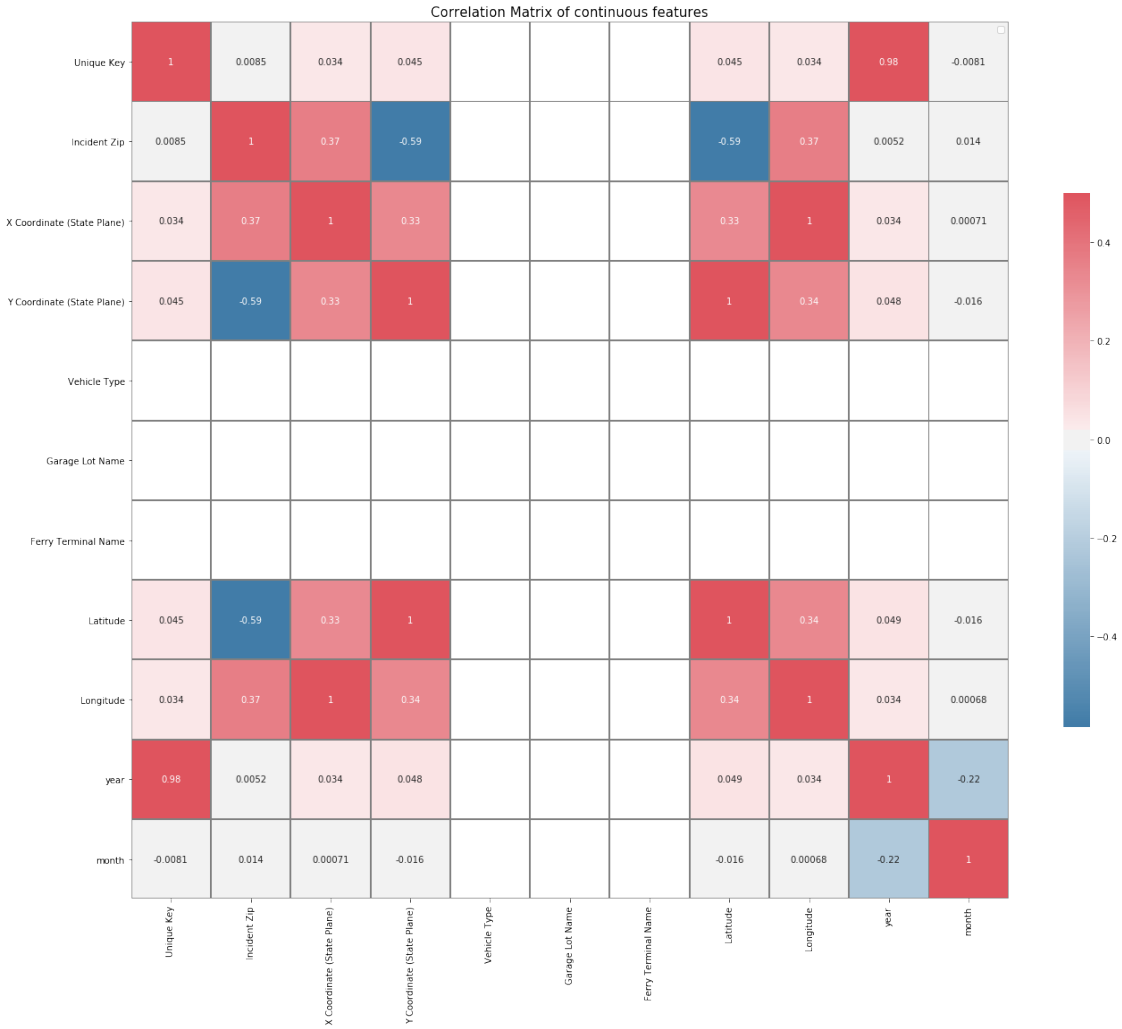












Runtime: 42 seconds