NYC BUSINESS DEVELOPMENT

CIS 9440 – FALL 2022

ERIC CHOW & JIAHONG YU

GOAL OF DATA WAREHOUSE / BACKGROUND

- Assist the New York City government in identifying the best location and business type to promote sustainable local economic growth
- Enable government agencies to better utilize public dollars on advertising, policing, and other economic incentives to attract customers to local businesses and develop the community

DATA SOURCES

- New York City Open Data – Legally Operating Businesses
  - https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh
- United States Census Data
  - https://data.census.gov/table

KPIS

- Number of new business per location
- Percentage of new businesses by industry
- Number of active businesses per population
- Active businesses per location
- Average length of business operation per industry

VISUALIZATION

- https://public.tableau.com/app/profile/jiahong.yu7692/viz/Milestone4_16693278940230/Dashboard2?publish=yes

UNIQUE CONSIDERATIONS – BUSINESS NAME

- There are two business name fields with inconsistent naming conventions
- Some records have both *business_name_2* and *business_name* populated
- And some with *business_name_2* populated have an individual's name populated in *business_name*

	business_name	address_city	address_state	address_zip	business_name_2
52	Recovery Racing V LLC	NEW YORK	NY	10019	Maserati of Manhattan
63	2474 QUAD VENTURES INC.	BROOKLYN	NY	11234	7 Eleven Store #35049A
64	IL RIFUGIO INC.	NEW YORK	NY	10024	TARALLUCCI E VINO
75	PYUN, KISUN	STATEN ISLAND	NY	10302	7-ELEVEN
81	STARBUCKS CORPORATION	BROOKLYN	NY	11209	STARBUCKS COFFEE COMPANY
84	MORGAN GLOBAL GROUP LLC	NEW YORK	NY	10019	CITY ZEN
88	JUAN, CARLOS FELIZ	NEW YORK	NY	10027	JC CENTER

UNIQUE CONSIDERATIONS – BUSINESS NAME (CONTINUED)

- We defined a function that would create a new field that we used as our de factor business name

```
# we'll create a business name column that would take the business_name_2 if the value is not null, oth
erwise it would take the value in business_name

# create a function that we'll then apply to each of the rows of the data frame for a newly defined 'ad
justed business name' column

def business_name_adj(df_row):
  # we initially used notnull to test whether cell value was null; however, the NaN were treated as flo
at and return an error
  # we instead converted the value for business_name_2 as string before testing whether it's a 'null' v
alue
  if str(df_row["business_name_2"]) != 'nan':
    return df_row["business_name_2"]
  else:
    return df_row["business_name"]
```

```
# we'll apply the previously created function to calculate a new adj_business_name column

data["adj_business_name"] = data.apply(business_name_adj, axis=1)
```

UNIQUE CONSIDERATIONS – STRFTIME METHOD

- We tried using the *strftime* method to convert Python date objects into a string ID in creating the date dimension
- However, there were values (e.g. NaT) that were invalid inputs for the *strftime* function
- We defined a function along with incorporating try/except to assign the correct date IDs

```
# upon further data profiling, we discovered that ~21k entries have NaT for license_expir_dd, which is
 not accepted by strftime
# we will create a function to set NaT to equal 99999999

def create_date_id(row, column_name):
  try:
    return pd.to_datetime(row[column_name]).strftime("%Y%m%d")
  except:
    return "99999999"
```

```
# create date ids for each for the license creation and license expiration columns

data["license_exp_date_id"] = data.apply(create_date_id, column_name='lic_expir_dd', axis=1)

data["license_creation_date_id"] = data.apply(create_date_id, column_name='license_creation_date', axis
=1)
```