Food Service Opportunities and Competition in the Kansas City Area: Data Sources and Python Code

Luther Nicholaus

Park University

Links to Data Sources

The two data sources for this project are:

- KCMO Food Permits
- Food Establishment Complaints

Data Preparation and Python Code

Data cleaning is at the heart of data preparation. The KCMO Food Permits data set is thoroughly cleaned to eliminate and correct irrelevant parts of the data, particularly in the "Establishment Name" column. In Python, irrelevant columns are dropped on both data sets. As shown in Fig. 1, there are many irrelevant parts to data defining Taco Bell establishments. The objective of data preparation is to have one stable format and name for all taco bell establishments. Failure to do so will result in inaccurate analysis and decision-making (data quality). Duplicate records, as shown in Fig. 2, are also removed for an accurate analysis.

The data set rows are also formatted to achieve consistency before analysis. Inconsistency in parts of data could eventuate into a contradictory analysis due to poor data quality. Failure to address inconsistency and irrelevant parts to the data could cost the entire project time, as the entire process ought to be repeated when subsequent issues transpire. Overall, the data preparation process, in particular data cleaning, consumes a sizeable chunk of time in this project. Gonzales (2004) notes "approximately 70 percent of data mining activity is focused on data preparation and cleaning" (p. 14).

Facility P...	Establishment Name
106790	Taco Bell #29844
11703	TACO BELL #2874
31432	TACO BELL
4479	Taco Bell
101161	Taco Bell #022580
101566	Taco Bell @ Tuileries Plaza Center
29981	Taco Bell #2953
11298	Taco Bell #4136
2600	TACO BELL #3266
102362	Taco Bell #21436
103809	Taco Bell #420 - Linwood
106763	Taco Bell #29842

Facility Permit #	Establishment Name
2600	Taco Bell
4479	Taco Bell
11298	Taco Bell
11703	Taco Bell
29981	Taco Bell
31432	Taco Bell
100216	Taco Bell
101161	Taco Bell
101566	Taco Bell
102362	Taco Bell
103809	Taco Bell
106762	Taco Bell
106763	Taco Bell
106790	Taco Bell
106799	Taco Bell
106973	Taco Bell

Fig 1. Removal of irrelevant parts of data.

e	Establishment Name	Facility Na	Facility Address
3	Aramark	Bartle Hall	301 W. 13th. St.
0	Aramark	Bartle Hall	301 W. 13th. St.
5	Aramark	Aramark a	601 E 12th St
9	Aramark	Royals Ara	1 Royal Way
0	Aramark	Royals Ara	1 Royal Way
0	Aramark	Royals Ara	1 Royal Way
2	Aramark	Royals Ara	1 Royal Way
4	Aramark	Royals Ara	1 Royal Way
5	Aramark	Royals Ara	1 Royal Way
6	Aramark	Royals Ara	1 Royal Way
7	Aramark	Royals Ara	1 Royal Way
8	Aramark	Royals Ara	1 Royal Way
2	Aramark	Chiefs Ara	1 Arrowhead Dr
1	Aramark	Chiefs Ara	1 Arrowhead Drive
4	Aramark	Chiefs Ara	1 Arrowhead Drive
5	Aramark	Chiefs Ara	1 Arrowhead Drive
6	Aramark	Chiefs Ara	1 Arrowhead Drive
8	Aramark	Chiefs Ara	1 Arrowhead Drive

	Facility Permit #	Establishment Name	Facility Address
435	100583	Aramark	301 W. 13th. St.
549	101495	Aramark	601 E 12th St
753	103969	Aramark	1 Royal Way

Fig 2. Removal of inconsistent parts of data.

```python
import pandas as pd
#from matplotlib import pyplot as plt
#from sklearn.cluster import KMeans

complaints = pd.read_csv('C:/Users/18166/Downloads/Food_Establishment_Complaints.csv')
food_permits = pd.read_csv('C:/Users/18166/Downloads/KCMO_Food_Permits.csv',encoding='latin1')


complaints.drop(['SOURCE', 'DEPARTMENT', 'WORK GROUP', 'CREATION MONTH', 'CREATION YEAR', 'STATUS', 'EXCEEDED EST TIMEFRAME', 'CLOSED DATE',
                 'CLOSED MONTH', 'CLOSED YEAR', 'DAYS TO CLOSE', 'ADDRESS WITH GEOCODE', 'NEIGHBORHOOD', 'COUNCIL DISTRICT', 'PARCEL ID NO',
                 'LATITUDE', 'LONGITUDE'], axis=1, inplace=True)
food_permits.drop(['Facility Name', 'Business Status', 'Facility Type', 'Permit Type', 'Web Site', 'Operational Status', 'Location 1'], axis=1, inplace=True)


food_permits = food_permits.drop_duplicates(subset=['Establishment Name', 'Facility Address'])

food_permits.to_csv('food_permits.csv')
complaints.to_csv('complaints.csv')

```

Fig 5. Code to remove redundant data and irrelevant columns.

References

Andrienko, N., Lammarsch, T., Andrienko, G., Fuchs, G., Keim, D., Miksch, S., & Rind, A. (2018). Viewing visual analytics as model building. *Computer Graphics Forum*, 37(6), 275-299. doi:http://dx.doi.org/10.1111/cgf.13324

Gonzales, M. L. (2004). The architecture of enterprise data quality. *Intelligent Enterprise*, 7(9), 14-17. Retrieved from https://search.proquest.com/trade-journals/architecture-enterprise-data-quality/docview/200663640/se-2?accountid=28370