Final Project: New York City Case Study using CRSIP-DM

Report 2: Association rule mining

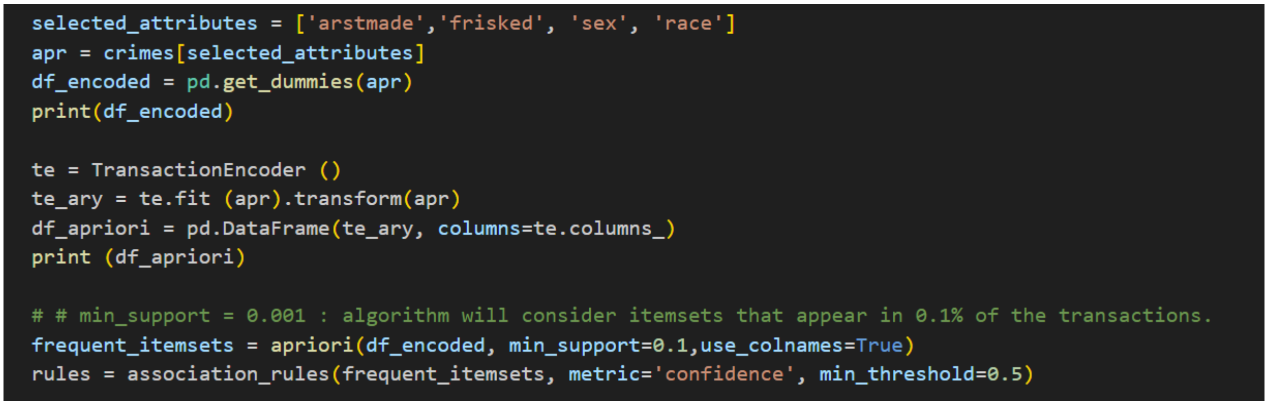
Juan J. Holguin

The Southern Alberta Institute of Technology

DATA 475: Advanced Concepts in Data Analytics

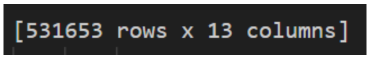
Data Preparation & modelling

For an APRIORI analysis, the dataset needs to be shortened to include just the required attributes. A “police force” data frame was processed without positive results. A “reason for search” data frame didn’t give accurate information as well. So the following data frame was prepared for a APRIORI model:



```
selected_attributes = ['arstmade','frisked', 'sex', 'race']
apr = crimes[selected_attributes]
df_encoded = pd.get_dummies(apr)
print(df_encoded)

te = TransactionEncoder ()
te_ary = te.fit (apr).transform(apr)
df_apriori = pd.DataFrame(te_ary, columns=te.columns_)
print (df_apriori)

# # min_support = 0.001 : algorithm will consider itemsets that appear in 0.1% of the transactions.
frequent_itemsets = apriori(df_encoded, min_support=0.1,use_colnames=True)
rules = association_rules(frequent_itemsets, metric='confidence', min_threshold=0.5)
```

The dummy variables dataset has the following shape.



```
[531653 rows x 13 columns]
```

Data evaluation

Frequent Itemsets

The most frequent itemsets are:

- (arstmade_N) with a support of 0.939301
- (frisked_N) with a support of 0.442108
- (frisked_Y) with a support of 0.557892
- (sex_M) with a support of 0.914021
- (race_B) with a support of 0.533239

Association Rules

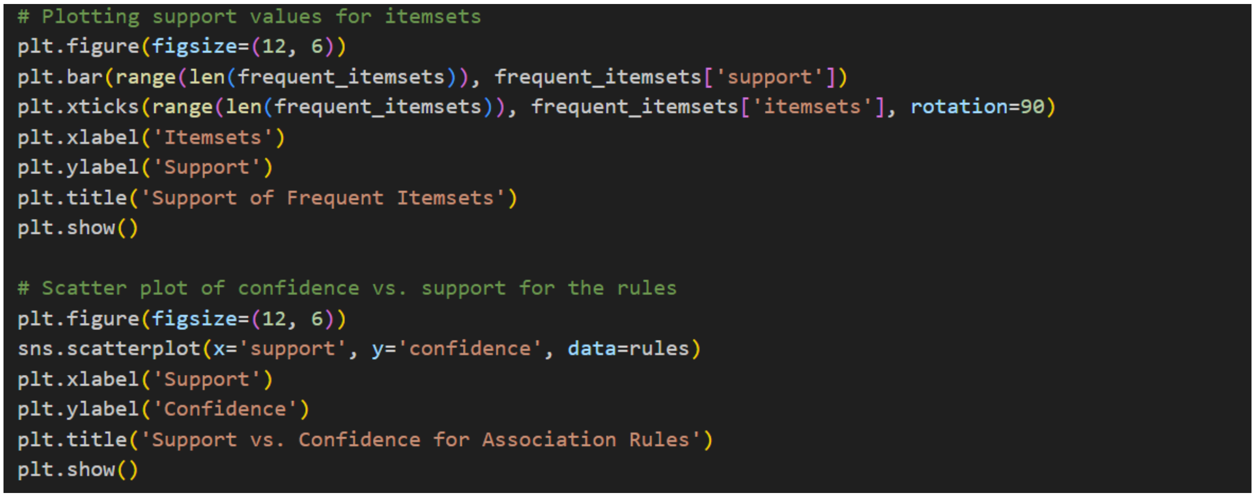
A selection of rules includes:

- **Rule**: (frisked_N) -> (arstmade_N), with support 0.432408, confidence 0.978060, lift 1.041264
- **Rule**: (sex_M) -> (arstmade_N), with support 0.860093, confidence 0.940999, lift 1.001808
- **Rule**: (arstmade_N) -> (frisked_Y), with support 0.506893, confidence 0.539649, lift 0.967300
- **Rule**: (race_B, frisked_Y) -> (arstmade_N), with support 0.306420, confidence 0.544455, lift 1.022259

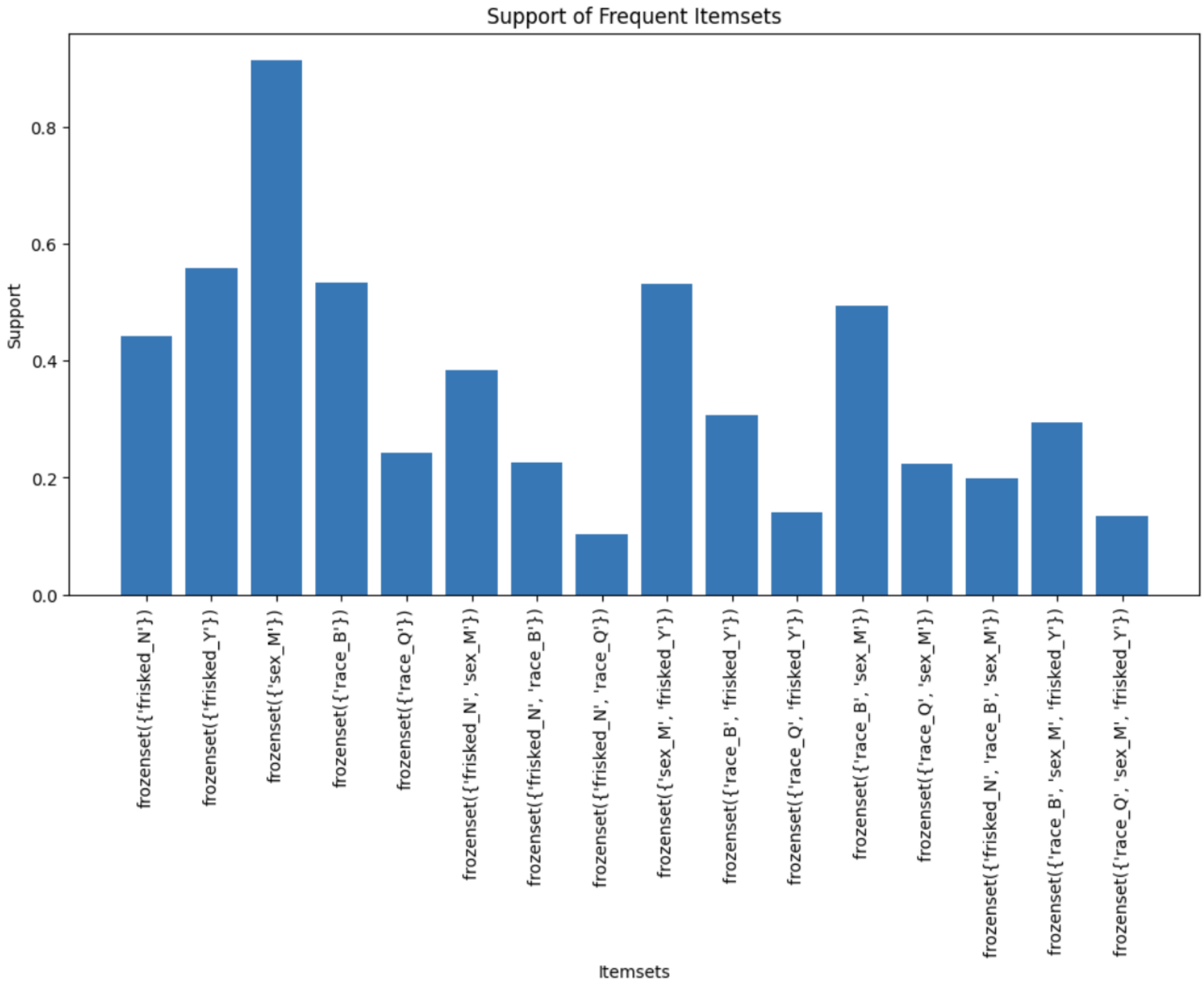
Key Metrics

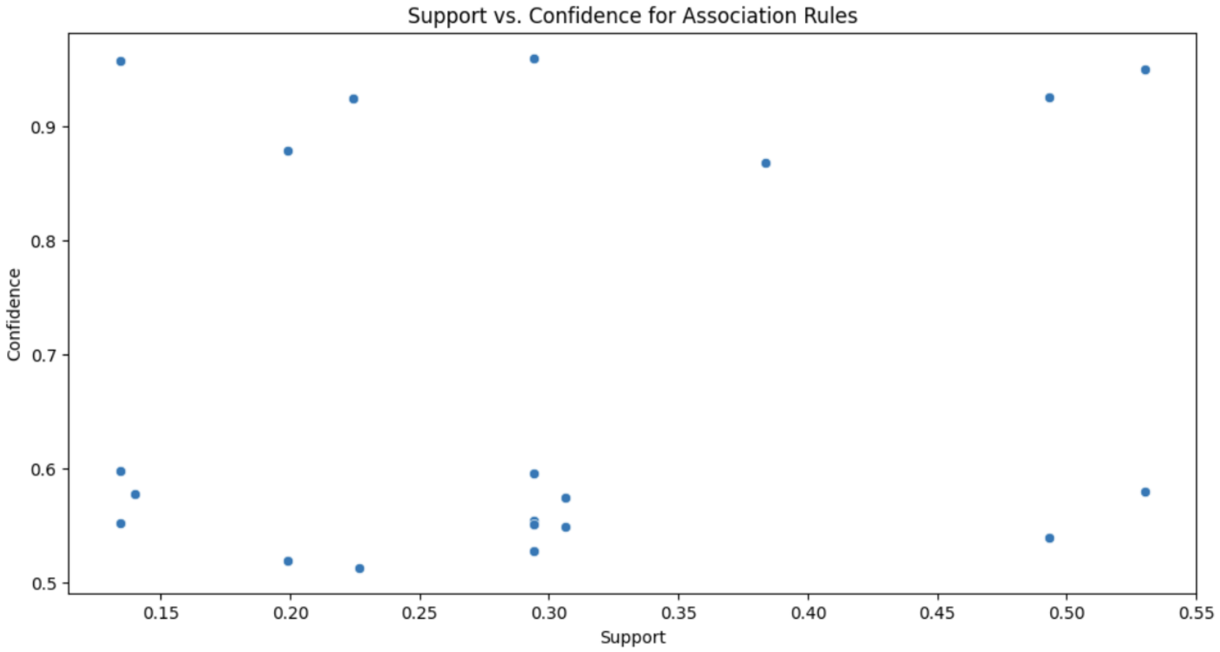
- **Support**: Frequency of itemsets/rules in the dataset.
- **Confidence**: Probability of the consequent given the antecedent.
- **Lift**: How much more likely the consequent is given the antecedent compared to random chance.
- **Leverage**: The difference between the observed frequency and expected frequency.
- **Conviction**: Measures the degree of implication of the antecedent on the consequent.
- **Zhang's Metric**: A variant of lift, indicating the strength of association.

The following charts will give a visual insight for the model:

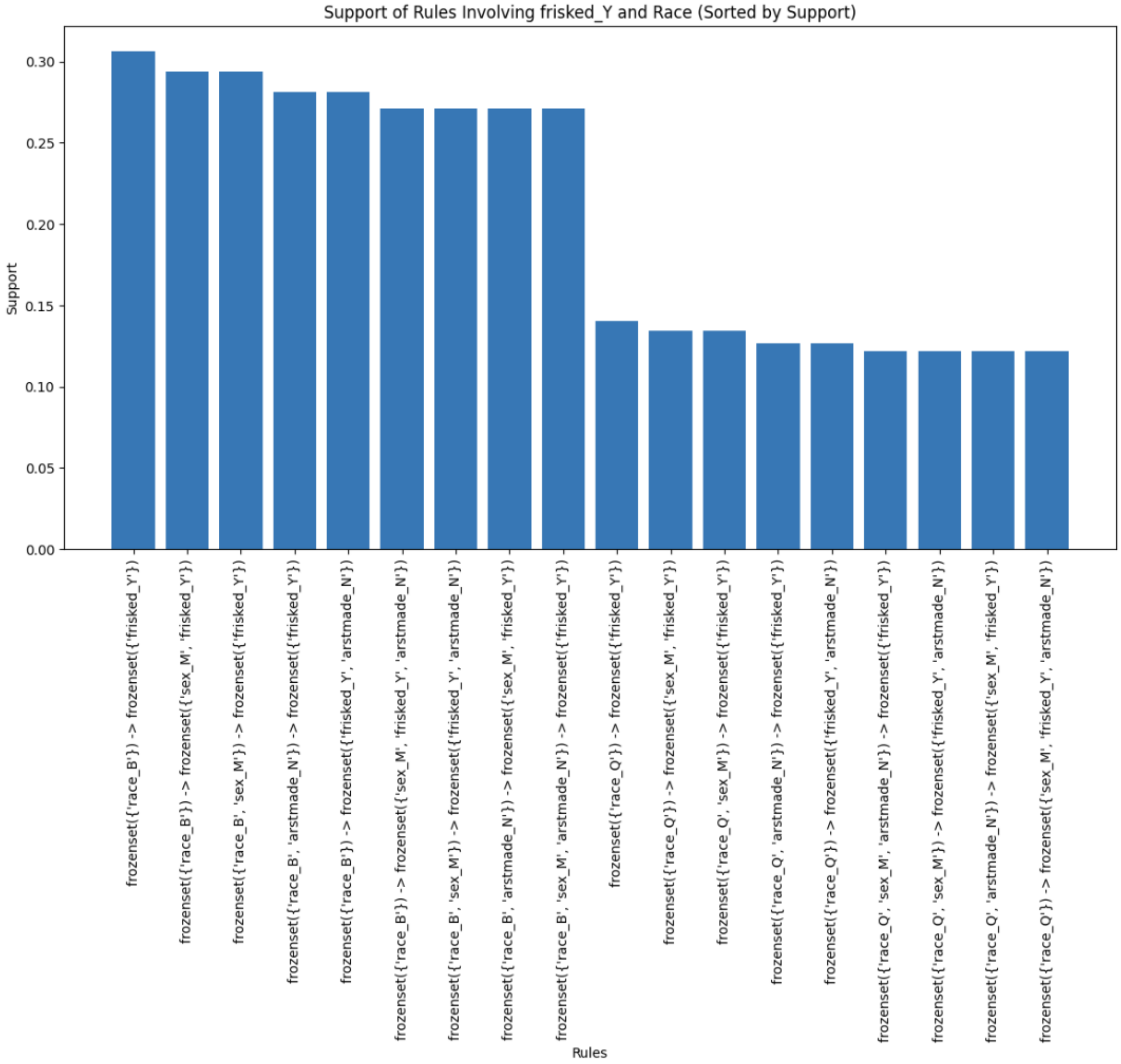


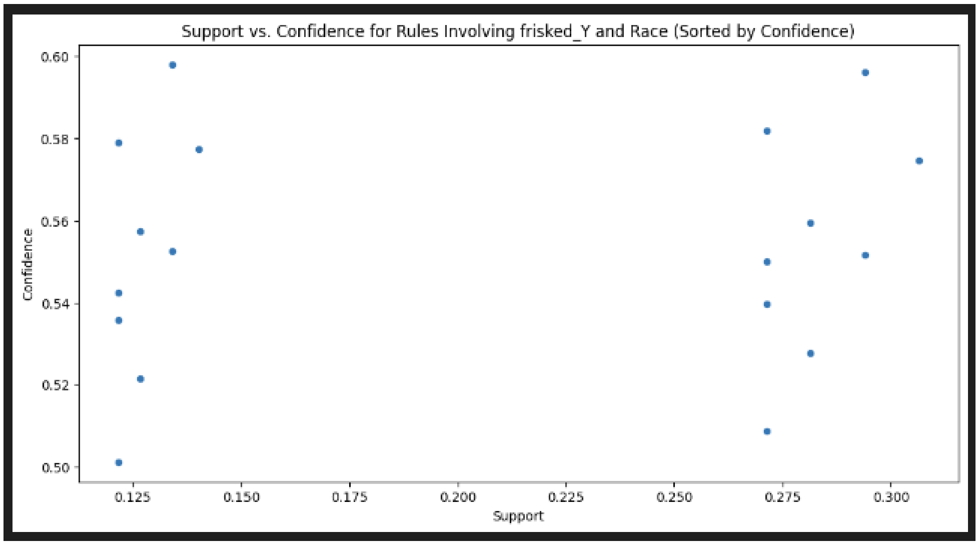
```python
# Plotting support values for itemsets
plt.figure(figsize=(12, 6))
plt.bar(range(len(frequent_itemsets)), frequent_itemsets['support'])
plt.xticks(range(len(frequent_itemsets)), frequent_itemsets['itemsets'], rotation=90)
plt.xlabel('Itemsets')
plt.ylabel('Support')
plt.title('Support of Frequent Itemsets')
plt.show()

# Scatter plot of confidence vs. support for the rules
plt.figure(figsize=(12, 6))
sns.scatterplot(x='support', y='confidence', data=rules)
plt.xlabel('Support')
plt.ylabel('Confidence')
plt.title('Support vs. Confidence for Association Rules')
plt.show()
```

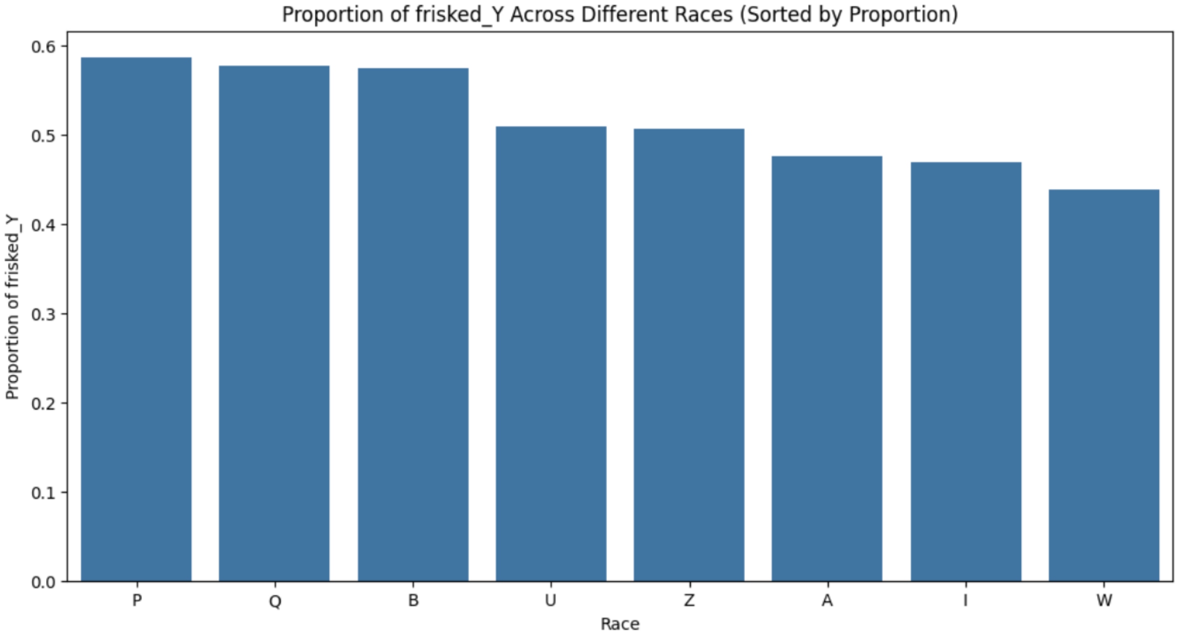

Support of Frequent Itemsets
Support
0.0
0.2
0.4
0.6
0.8
frozenset({'frisked_N'})
frozenset({'frisked_Y'})
frozenset({'sex_M'})
frozenset({'race_B'})
frozenset({'race_Q'})
frozenset({'frisked_N', 'sex_M'})
frozenset({'frisked_N', 'race_B'})
frozenset({'frisked_N', 'race_Q'})
frozenset({'sex_M', 'frisked_Y'})
frozenset({'race_B', 'frisked_Y'})
frozenset({'race_Q', 'frisked_Y'})
frozenset({'race_B', 'sex_M'})
frozenset({'race_Q', 'sex_M'})
frozenset({'frisked_N', 'race_B', 'sex_M'})
frozenset({'race_B', 'sex_M', 'frisked_Y'})
frozenset({'race_Q', 'sex_M', 'frisked_Y'})
Itemsets


Support vs. Confidence for Association Rules
Confidence
0.5
0.6
0.7
0.8
0.9
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
Support

Furthermore, if we analyze frisked_Y and races we get the following results:






Proportion of frisked_Y Across Different Races (Sorted by Proportion)
Proportion of frisked_Y
0.0
0.1
0.2
0.3
0.4
0.5
0.6
P
Q
B
U
Z
A
I
W
Race