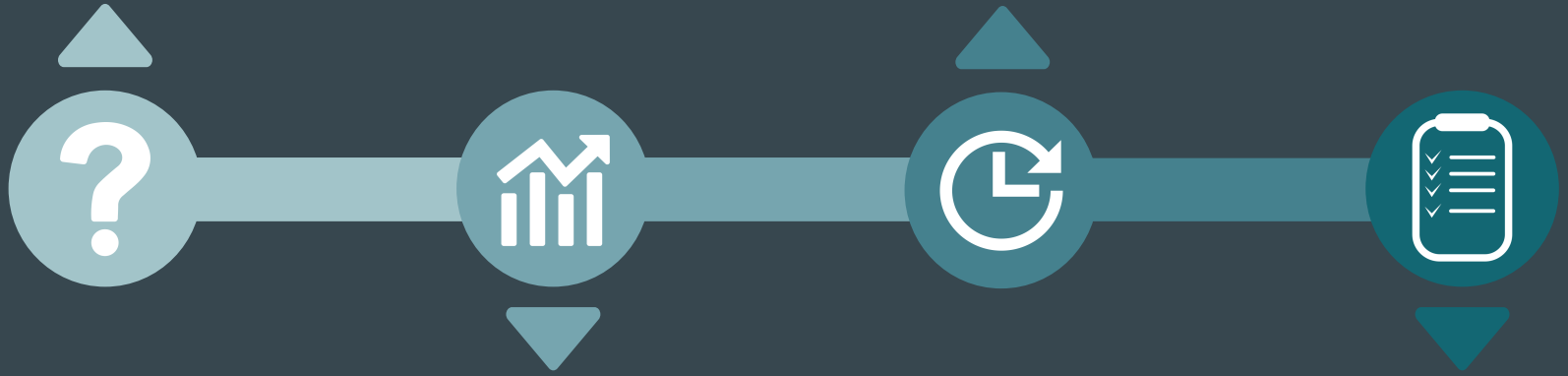
Exploring Gas Station Expansion with Analytics

Vignesh Mulay, Celina Paudel, Benjamin McGriff

Today's Agenda

Empire State Gas & Fuel

Problem and Business Objectives



Predictive Analytics (NYC Census Data)

Recommendations and Next Steps

Final Conclusions

Business Problem

BACKGROUND

Empire State Gas & Fuel is an American chain of gas stations based in the state of New York.

The company operates throughout the state but has no filling stations in New York City.

OBJECTIVES

Our main goal is to increase our share of the New York gas market through expansion.

We want to open our first gas station in New York City due to it being so densely populated.

PROBLEM

"In which area of NYC should we open a new gas station in order to reach the most customers?"

We aim to use analytics to determine the best location for a station based on 3 key factors.

Indicators of a Strong Location

Empire State
Gas & Fuel

Multiple Linear Regression: Commute Time

Built a multivariate linear regression model that predicts mean commute time

Used stepwise selection to remove any variables that were not relevant to the model

Has a relatively high correlation accuracy of 76.4% which suggests that the actual and predicted values move in a similar direction

15 predictor variables were found to be statistically significant

Highest Relative Variable Importance:

1. County: 18.4%
2. SelfEmployed: 8.91%
3. PrivateWork: 7.36%
4. Service: 5.29%

Occupation type seems to be highly influential with a higher percentage of service workers being the largest positive driver

Kings County (Brooklyn) tends to positively affect commute time; the opposite is true for New York County (Manhattan).

Multiple Linear Regression: Driver/Carpool %

Empire State Gas & Fuel

Built a multivariate linear regression model that predicts Driver % based on various factors

Once again, stepwise selection was used to remove any irrelevant variables

Has high correlation accuracy of 85.4%

Once again, a surprisingly large total of 16 variables were statistically significant

Highest Relative Variable Importance:

1. Private Work: 17.5%
2. White: 17.1%
3. Office: 12.9%
4. Asian: 12.1%

While occupation is relevant, race seems to be particularly important when predicting Driver %

Richmond County (Staten Island) is the only county that positively affects Driver %

Binary Logistic Regression: Household Income Level

Empire State Gas & Fuel

Built a logistic regression model to predict whether a census tract has “Above Average” or “Below Average” Household Income

Used sociodemographic factors including race, occupation, and residency information as predictor variables

Model seems to be highly effective as it has:

- ➔ High prediction accuracy: 85%
- ➔ Excellent classification accuracy: 0.934

Significant Influencing Variables:

1. Driver/Carpool %
2. Borough (Queens)
3. Commute Time
4. Asian (%)

Significance of Driver/Carpool % and Commute Time supports our location strategy

Queens has the largest positive effect out of all variables indicating that it is a potentially rewarding location option

Future Recommendations and Next Steps

Empire State Gas & Fuel

LIMITATIONS

Linear regression assumes data is entirely independent

Linear regression focuses on the mean of the DV and not its extremes

Logistic regression is only truly useful if all the relevant IVs are identified

Data like real estate costs and current gas station locations is still needed

RECOMMENDATIONS

Use clustering to identify groups of census tracts with similar traits

Use quantile regression to model the 90th percentile of the outcome variable

Collect and include education and industry data to the logistic model

Access other databases to find the following information

Conclusion

Empire State Gas & Fuel

Problem and Objective → **Analytics and Modelling** → **Value Added**

Empire State Gas & Fuel wants to open a gas station in NYC

We want to determine the best area in NYC to open a gas station

Linear Regression:
Commute Time

Linear Regression:
Driver/Carpool %

Logistic Regression:
Household Income Level

Best Location:

- Queens, Brooklyn, or Staten Island
- High % employed in service jobs
- Low % self-employed or employed in private industry

Questions?