

Project 4 – Greece Travel Insights

By: Reinforcement Learning Rockstars

Mavin Gill, Natalia Lopez, Andrew
Rexford and David Kauffman



Introduction

- Retrieved our dataset from Kaggle
- Project focused on analyzing Greek tourism insights
- Some of our topics of discovery were:
 - Best times to travel to Greece
 - What type of accommodation people used when travelling
 - Exploring relationships between cost and accommodation, date of travel and length of stay

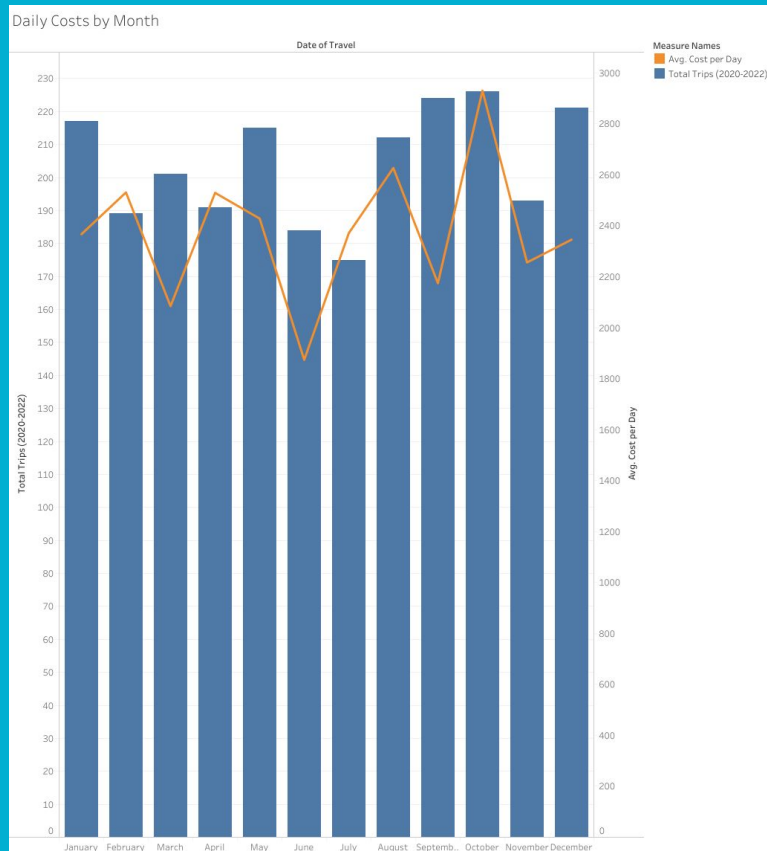


Data Source and Limitations

- Greek tourism data from Kaggle
 - Data from January 2020 - August 2023
 - 3,000 rows, 13 columns
- Mix of authentic and synthetic data.
- Data source did not specify a currency in the documentation
 - We assumed cost in Euros
- Cost column varies wildly
 - Examples:
 - 28 day trip costing 5,032
 - 1 day trip costing 29,909

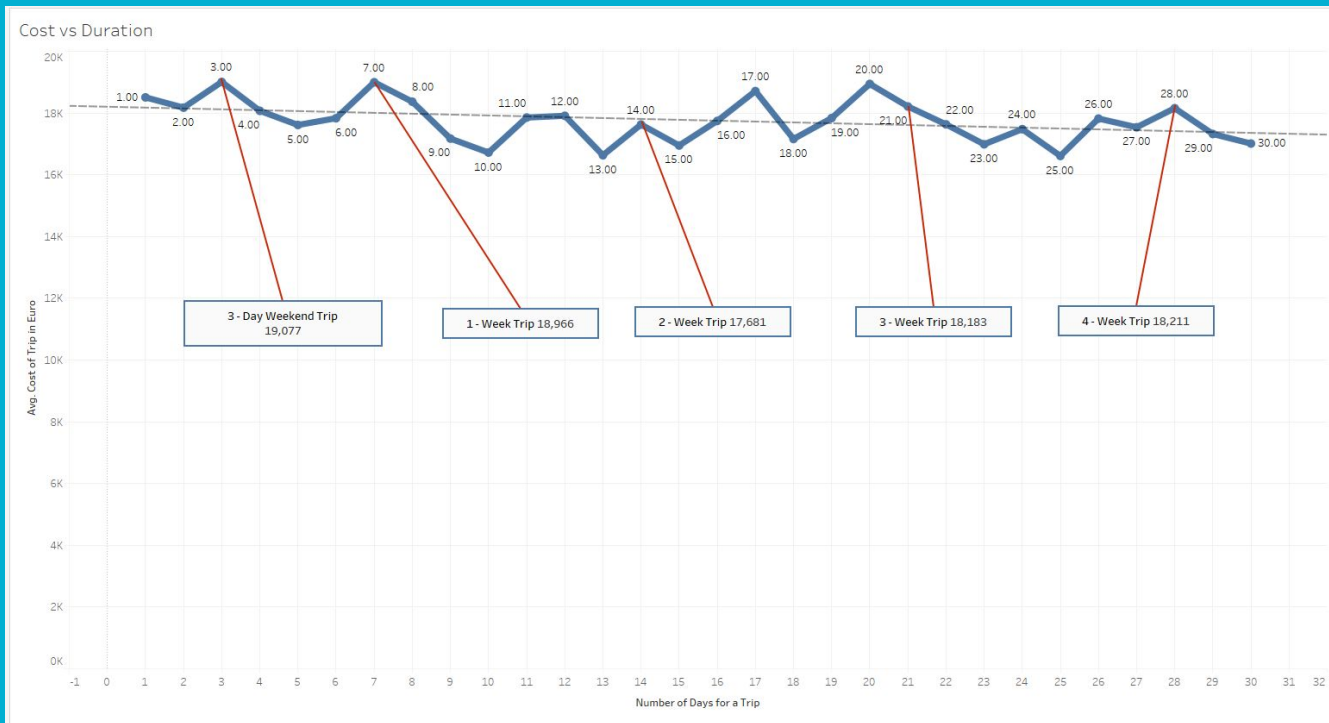
What are the best months to travel to Greece?

- Most Affordable:
 - June
 - March
 - September
- Smallest Crowds:
 - July
 - June
 - February
- Best Overall: June
- Worst Overall: October



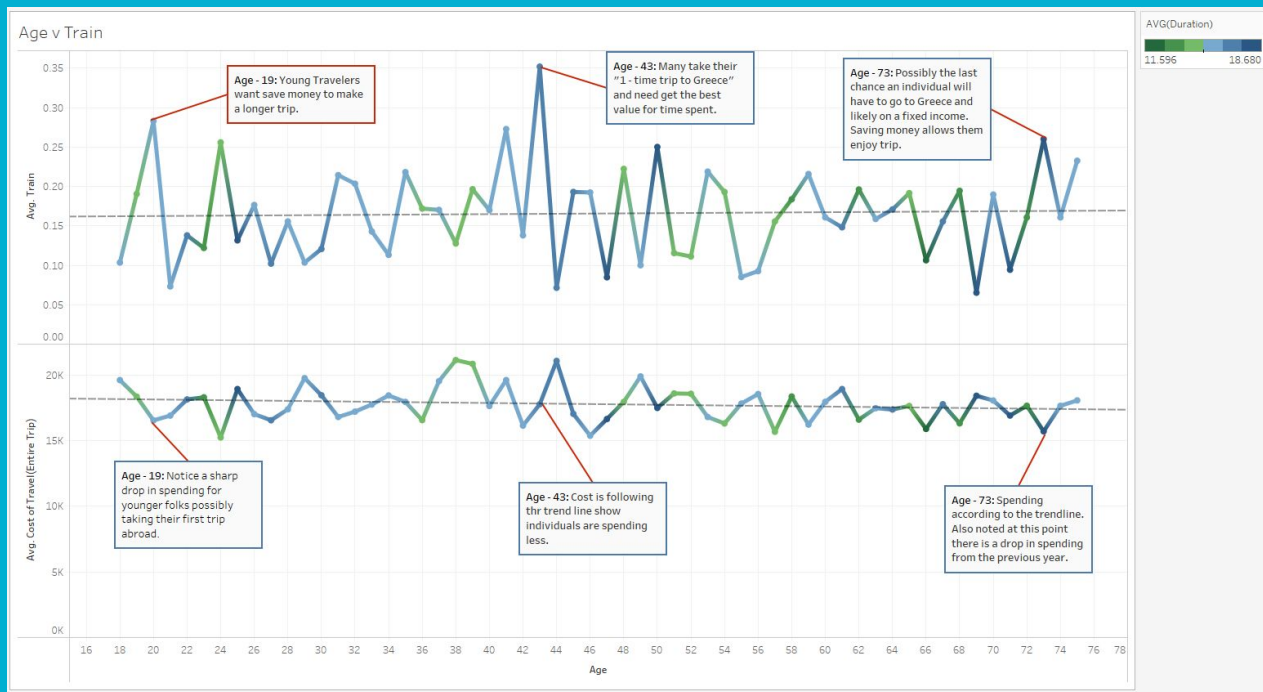
Andrew Insights

Cost of a Trip vs Number of Days for a Trip



Andrew Insights

Age vs Traveling by Train with Respect to Duration of Trip



Machine Learning

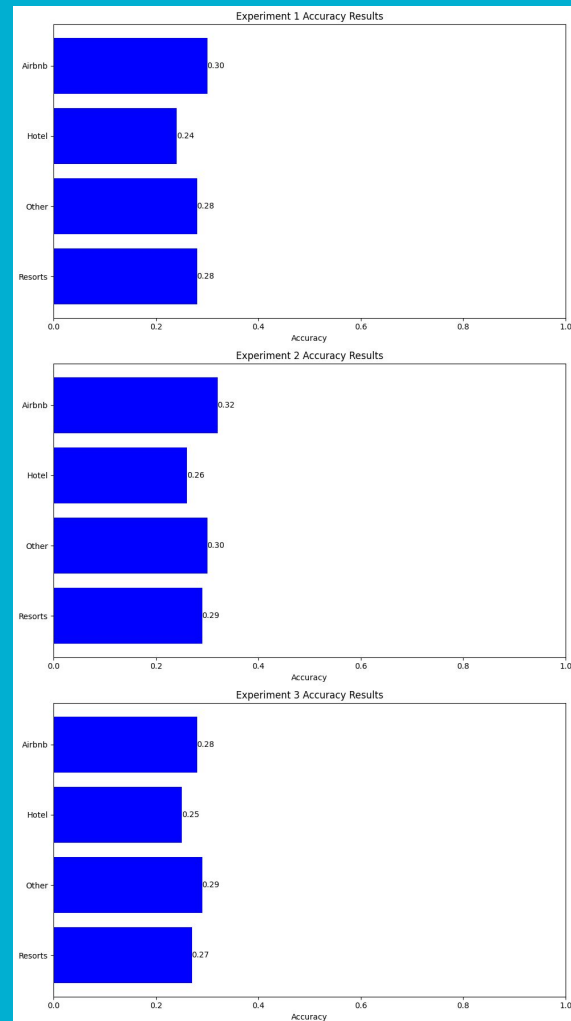
Type of Accommodation:

- We created a Logistic Regression model focused on the 'type of stay' that tourists used most often.
- Ran 3 models:
 - 500 iterations
 - 750 iterations
 - 1,000 iterations
- Accuracy of the model with 1,000 iterations was the strongest, at 27%



Machine Learning – Visuals

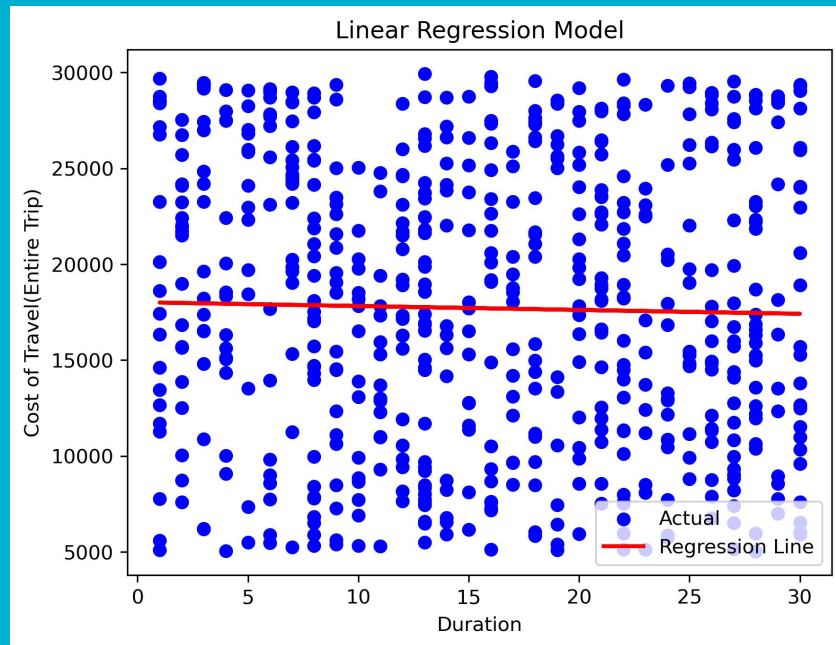
- Using Matplotlib, we created visual representations of the 3 trials.
- Airbnb had the best overall accuracy scores, with an average accuracy percentage of 30% between the 3 models.



Machine Learning (2)

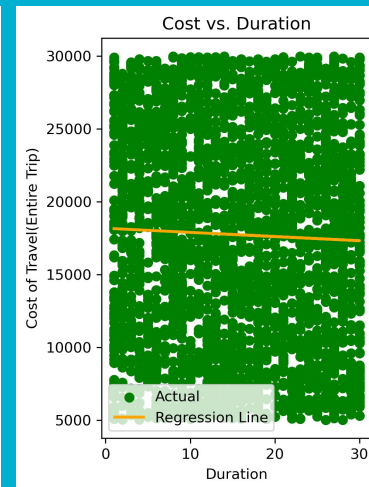
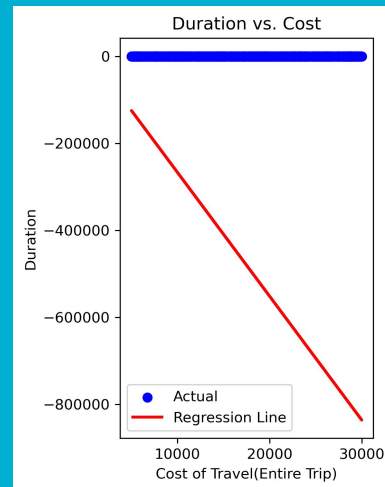
Duration of the Stay and Total Cost.

- We created a Logistic Regression model focused on the 'Duration of the Stay'.
- Ran several models. Accuracy of the model:
 - Default: 38%
 - 500 iterations: 55 %
 - 750 iterations: 55%
 - 1,000 iterations: 62%
 - 2,000 iterations: 79%
- Accuracy of the model with 2,000 iterations was the strongest, at 79%



Machine Learning (2)

- 2 Regression models to predict:
 - Duration of Stay based on Cost
 - Cost based on Duration of Stay
 - Input: Number of days
 - Output: Cost



Limitation:

- Uniform dataset due to synthetic data.
- Problems when making predictions.

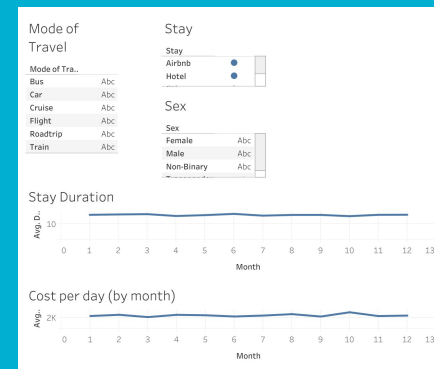
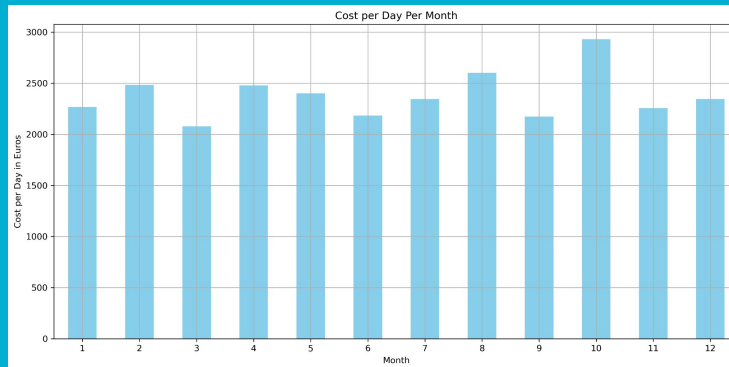
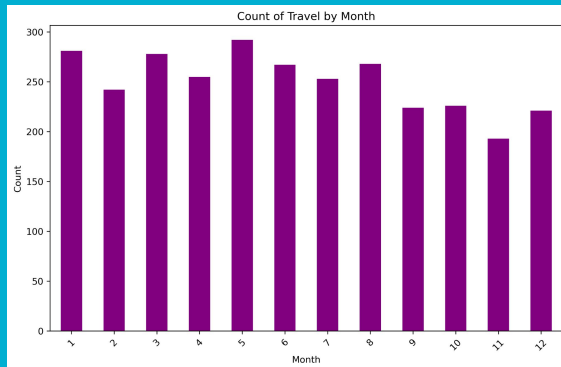
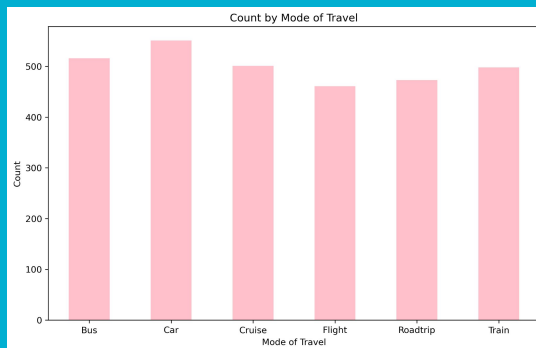
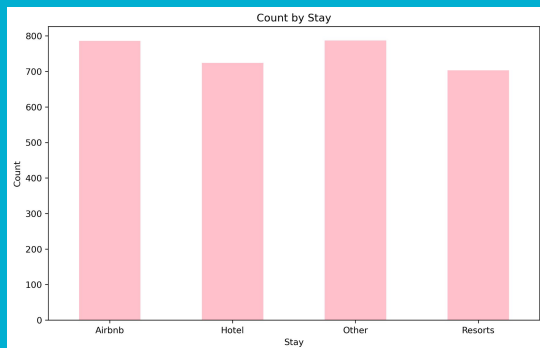
```
# Predict the cost for a 25-day stay
predicted_cost = cost_model.predict(duration_input)

print(f"Predicted Cost for a 25-day stay: ${predicted_cost[0]:.2f}")

Predicted Cost for a 25-day stay: $17475.93
Predicted Cost for a 10-day stay: $17903.16
```

ML Insights

Exploring the dataset and finding limitations due to synthetic data.

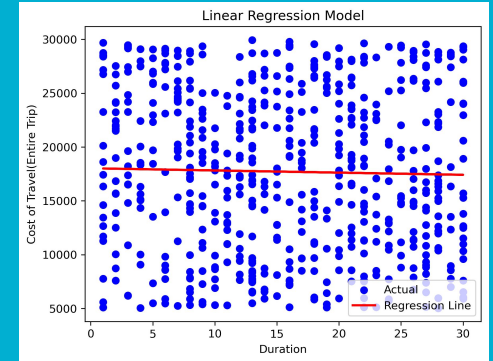


Conclusion

- Best month to travel: June
- Dataset was not very conducive to accurate machine learning models.
 - May be due to inclusion of synthetic data

Our Analysis shows a Linear Regression Model is not optimal for the following reasons:

- The Variance is too high, data is perfectly space with very little clustering, not enough data close to the mean.
- The R^2 (coefficient of determination) is incredibly low.



Thank You!

