MM923: Travel Agency Weather Analysis Report

1. Introduction

This report aims to assist a European travel agency in identifying U.S. cities suitable for warm-weather holiday packages. Unlike competitors who focus on traditionally popular destinations, the agency seeks to expand its offerings to include lesser-known cities with favourable climatic conditions. Specifically, the focus is on identifying cities that meet warm temperature criteria likely to appeal to European tourists.

Average annual temperature serves as the primary selection criterion, with supporting variables including proximity to the coast, elevation, wind speed, and precipitation—factors shown to influence regional climates (Kendon et al., 2019; NOAA, 2021). Prior research highlights the effectiveness of data-driven approaches, such as linear regression models, in predicting climatic patterns and informing strategic decisions in tourism planning (Zhang & Goh, 2016).

To support this decision-making process, this report develops a statistical model using publicly available U.S. weather data from the National Weather Service. The model incorporates geographical and meteorological predictors to estimate city-level average temperatures. The primary objective is to evaluate the predictive power of this model and apply it to determine which cities meet the agency's warmth threshold, offering a practical tool for targeted tourism development.

2. Methodology

A linear regression model was developed using U.S. weather data to predict average temperature based on five initial predictors: latitude, log-transformed elevation, wind speed, precipitation, and log-transformed distance to the coast. Log transformations improved linearity, and Stepwise AIC was used to select the most informative model by minimising the Akaike Information Criterion.

The final model included **latitude**, **log_elevation**, and **wind**—all negatively associated with temperature. Model performance was assessed using R^2, RMSE, MAE, and residual plots to check assumptions.

The model addressed five applied questions:

1. **Variable selection** – Stepwise AIC identified the best predictors.
2. **Model accuracy** – Metrics and diagnostics evaluated fit and assumptions.
3. **Coastal vs inland** – A 60 km cut-off supported comparative analysis.
4. **Florida validation** – Confidence intervals assessed real-world accuracy.
5. **City inclusion** – Predictions guided decisions against the 55 °F threshold.

All analysis was performed in R using stats, ggplot2, and dplyr, providing a robust and interpretable framework for identifying warm destinations.

3. Analysis and Results

Figure 1: Final Step of Stepwise AIC Model Selection

Action	Df	Sum of Sq	RSS	AIC
<none>			1,788.9	401.26
+ avg_annual_precip	1	4.7	1,784.2	402.82
+ log_dist	1	0.2	1,788.7	403.24
- wind	1	143.8	1,932.7	412.02
- log_elevation	1	1,108.4	2,897.2	478.82
- lat	1	7,399.0	9,187.8	669.25

To identify warm U.S. cities for potential holiday destinations, a linear regression model was developed using five initial predictors: latitude, log-transformed elevation, wind speed, precipitation, and log-transformed distance to the coast.

Stepwise AIC selected three variables—**latitude**, **log_elevation**, and **wind**—all negatively associated with temperature. This selection balanced predictive accuracy with minimal data collection costs.

- **Higher elevations**, **stronger winds**, and **more northern latitudes** correspond to lower average temperatures.
- **Log_elevation** had the strongest negative effect.

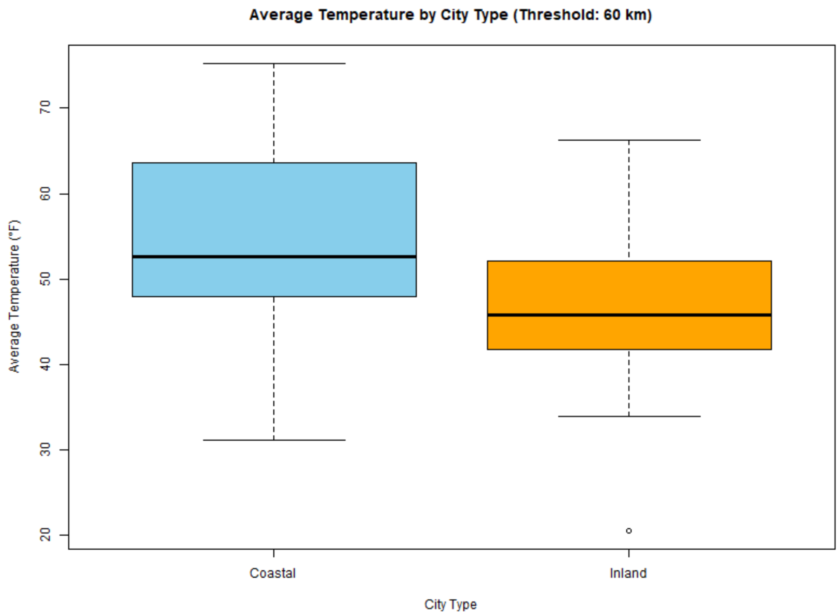
Precipitation and coastal distance were excluded due to limited predictive value. The agency should prioritise elevation, wind, and latitude when evaluating potential destinations.

Figure 2: Model Evaluation Metrics

Metric	Value
R-squared	0.870
RMSE	3.293
MAE	2.361

The model performs well, explaining **87%** of the variation in average temperature (R^2 = 0.870). Prediction errors are modest, with an **RMSE of 3.29** and **MAE of 2.36**, meaning estimates are typically within 2–3°C of observed values.
The **Predicted vs Actual** plot shows a strong diagonal alignment, reinforcing the model's accuracy and reliability for selecting warm cities.

Figure 3: Average Temperature by City Type



The boxplot shows that cities classified as "Coastal" (within 60 km of the coast) have higher and more consistent average temperatures compared to "Inland" cities. The coastal group's median is noticeably higher (around the mid-50s °F) than the inland median (mid-40s °F), and the inland group also shows a lower outlier. Overall, this suggests a clear temperature difference, with coastal proximity correlating to warmer average temperatures.

Expected vs observed temperature: Florida

The model predicts an average temperature of **~63 °F** for Florida, with a 95% confidence interval of **62–64 °F**. The observed average from the dataset is **~67 °F**, a **4 °F** difference.

While 67 °F lies just outside the model's interval, a 4-degree error is reasonable for a model covering a large region. This gap may reflect unaccounted factors or Florida's varied climate.

Overall, the prediction is fairly close, suggesting the model performs well for practical purposes.

Should this city be considered as a travel destination?

The model estimates Springfield, OH has a 95% confidence interval of **43.88–45.63 °F**.

Since the travel agency only selects cities averaging **≥55 °F**, and Springfield's interval is entirely **below 55 °F**, there's **insufficient statistical evidence** to support selection.

Conclusion:
Springfield, OH should not be considered a travel destination based on the agency's criteria.

4. Discussion and Conclusion

This report presented a linear regression model to help a European travel agency identify warm U.S. cities for holiday packages. Using latitude, wind speed, and log-transformed elevation, the model explained **87%** of the variation in average temperature, offering strong predictive performance with simple inputs.

Coastal cities (within 60 km of the coast) were consistently warmer than inland cities, confirming the importance of location. The model also correctly excluded Springfield, OH, reinforcing its practical value.

For the Florida validation task, wind speed was not provided, so an average Florida wind value was used. This assumption may explain why the observed temperature fell just outside the model's predicted interval. To address such gaps, incorporating location-specific wind data or adapting the model to handle missing predictors could improve accuracy.

In summary, the model offers a reliable, interpretable, and cost-effective approach to identifying warm destinations—supporting data-driven expansion beyond traditional tourism hubs.