STAT 425 Graduate Project Final Report

Due Tuesday, December 16, 11:59 pm.

Submit through Canvas.

Name: Ting Yun Chen

Netid: tingyun3

Organize your project report in this document and submit the qmd file, the rendered pdf file, and any data or image files needed to render the report.

The final report counts for 60 points out of the 80 points available for the project as a whole. You can delete the instructions before submitting your report.

Summary and motivation (1-2 paragraphs, 12 pts)

As an international student, the desire to travel home to visit family and friends is often constrained by the high cost of international flights, especially during peak travel seasons. Airline ticket prices are influenced by various factors—including route, airline, cabin class, departure and arrival times, flight duration, and booking lead time—yet the combined effects of these determinants in dynamic pricing systems remain opaque to consumers. Prior work in airline economics shows that carriers systematically adjust fares over time through intertemporal price discrimination and dynamic responses to uncertain demand and remaining seat capacity (Williams, 2018), motivating a closer empirical look at how observable flight characteristics and purchase timing translate into price differences.

The purpose of this project is to use statistical modeling techniques covered in STAT-425 to predict airline ticket prices and identify which factors most strongly drive fare variation. I compare multiple modeling strategies, including multiple linear regression (MLR), model selection methods, and regularization techniques such as Ridge and Lasso regression, to evaluate

the trade-off between predictive accuracy and interpretability. The Lasso, in particular, is designed to perform simultaneous shrinkage and variable selection by constraining the L1 norm of the coefficients (Tibshirani, 1996), while modern algorithms such as glmnet compute entire regularization paths efficiently via coordinate descent (Friedman, Hastie, & Tibshirani, 2010). These tools are well-suited for the high-dimensional dummy encodings required by airline and route factors in this dataset.

This project uses the publicly available *Flight Price Prediction* dataset originally released on Kaggle by the Easemytrip team. The dataset contains approximately 300,000 flights between six major Indian metropolitan cities. Dataset URL: https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

Methods (1-3 paragraphs, 12 pts)

I modeled ticket price as a continuous response and first fit a baseline multiple linear regression (MLR) including all available predictors. Because the dataset contains several categorical variables (airline, origin, destination, cabin class, number of stops, and departure/arrival time bins), the baseline model has an ANCOVA structure with dummy variables.

Model diagnostics were used to assess linearity, homoscedasticity, and residual normality. To address nonlinear relationships between price and the continuous predictors (`duration` and `days_left`), I then fit an extended model that replaces these linear terms with B-spline basis expansions with four degrees of freedom, allowing for flexible shapes in the duration–price and lead-time–price relationships that are consistent with dynamic pricing behavior documented in the airline literature (Williams, 2018).

To improve out-of-sample predictive performance and study variable importance, I applied regularization using the `glmnet` package. Lasso regression ($\alpha = 1$) was used for variable selection, leveraging the L1 penalty's ability to shrink many coefficients exactly to zero and produce sparse, interpretable models (Tibshirani, 1996). Ridge regression ($\alpha = 0$) was used to stabilize coefficients in the presence of many correlated dummy variables through L2 shrinkage. Both models were estimated along a regularization path using cyclical coordinate descent (Friedman et al., 2010) and tuned via 5-fold cross-validation, with root mean squared error (RMSE) as the primary performance metric.

In addition to regression, I used one-way and two-way ANOVA to formally test price differences across cabin classes and airlines, including an airline $\times$ class interaction. Finally, I fit a linear mixed-effects model with random intercepts for airlines and for origin–destination routes to capture unobserved heterogeneity in baseline price levels. All models were implemented in R using the `stats`, `splines`, `glmnet`, `car`, and `lme4` packages.

Data (12 pts)

```
library(readr)
library(tidyverse, quietly = TRUE)
library(car)
library(splines)
library(glmnet, quietly = TRUE)
library(broom)
library(lme4)
library(doParallel)
registerDoParallel(4)

flight_data <- read_csv("Clean_Dataset.csv")

# remove index column
flight_data <- flight_data %>%
  select(-`...1`) %>%
  mutate(
    across(
      c(airline, flight, source_city, destination_city,
        class, stops, departure_time, arrival_time),
      as.factor
    )
  )

summary(flight_data)
```

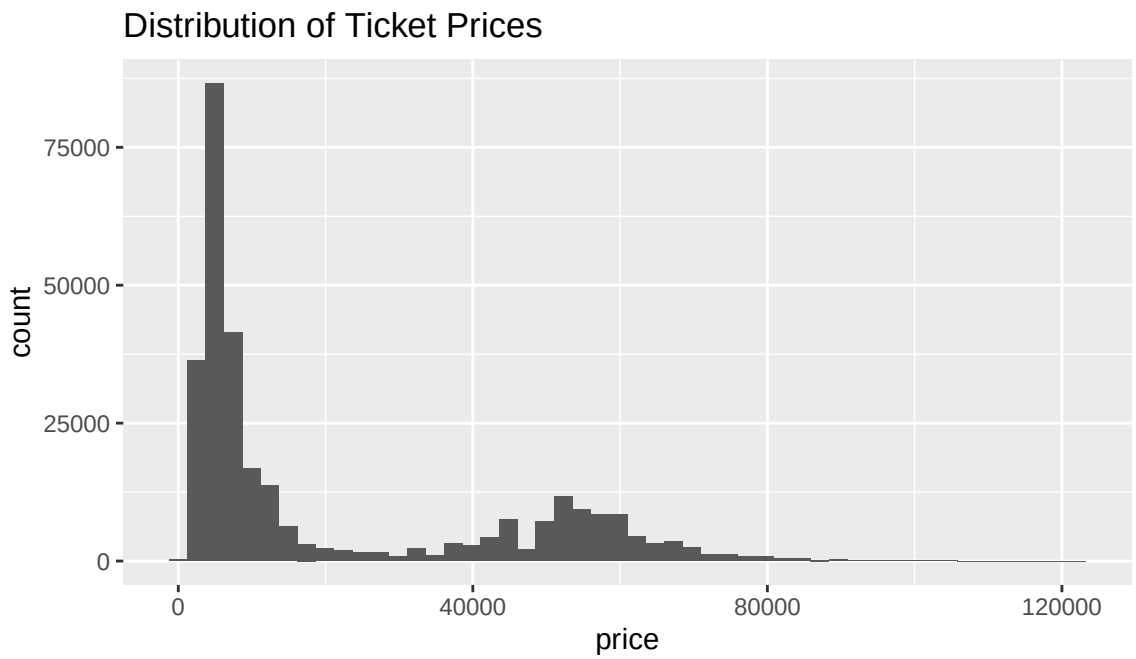
```
      airline           flight          source_city          departure_time
 Air_India: 80892   UK-706 :  3235   Bangalore:52061   Afternoon    :47794
 AirAsia  : 16098   UK-772 :  2741   Chennai  :38700   Early_Morning:66790
 GO_FIRST : 23173   UK-720 :  2650   Delhi    :61343   Evening      :65102
 Indigo   : 43120   UK-836 :  2542   Hyderabad:40806   Late_Night   : 1306
 SpiceJet :  9011   UK-822 :  2468   Kolkata  :46347   Morning      :71146
 Vistara  :127859   UK-828 :  2440   Mumbai   :60896   Night        :48015
                    (Other):284077
        stops              arrival_time    destination_city      class
 one        :250863   Afternoon    :38139   Bangalore:51068   Business: 93487
 two_or_more: 13286   Early_Morning:15417   Chennai  :40368   Economy :206666
 zero       : 36004   Evening      :78323   Delhi    :57360
                      Late_Night   :14001   Hyderabad:42726
                      Morning      :62735   Kolkata  :49534
```

```
                    Night        :91538   Mumbai   :59097

    duration       days_left      price
 Min.   : 0.83   Min.   : 1   Min.   :  1105
 1st Qu.: 6.83   1st Qu.:15   1st Qu.:  4783
 Median :11.25   Median :26   Median :  7425
 Mean   :12.22   Mean   :26   Mean   : 20890
 3rd Qu.:16.17   3rd Qu.:38   3rd Qu.: 42521
 Max.   :49.83   Max.   :49   Max.   :123071
```

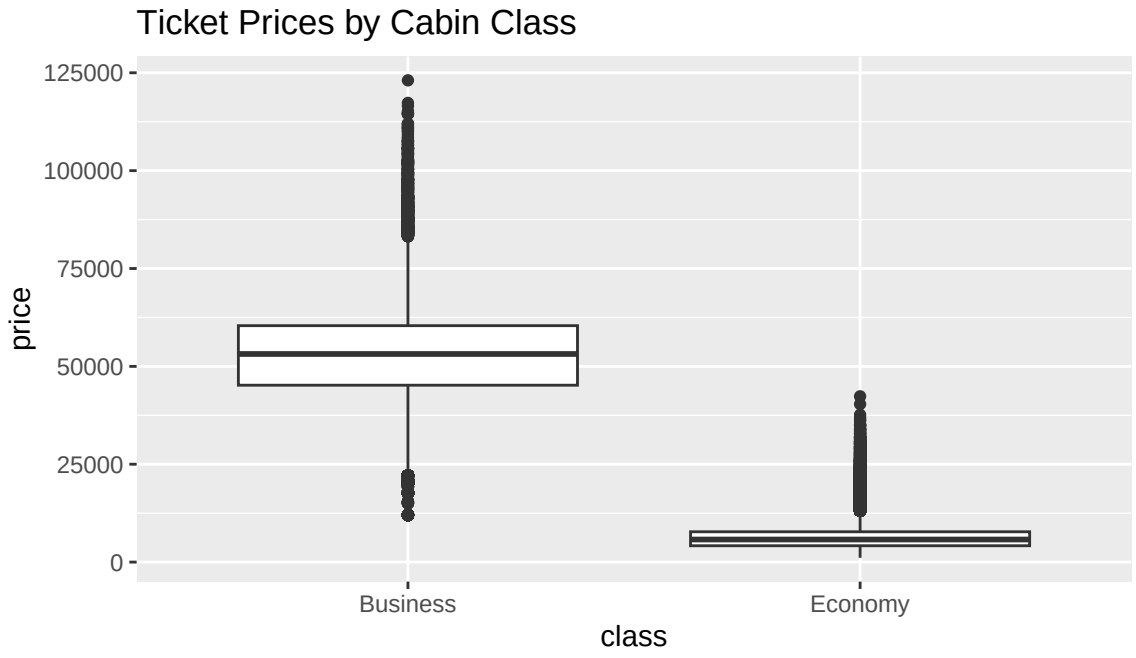
The cleaned dataset contains 300,153 flights and 11 variables after removing an index column. Each row corresponds to a scheduled commercial flight between major Indian metro cities. Categorical variables describe the airline, origin and destination city, cabin class, number of stops, and departure/arrival time-of-day, while continuous variables include flight duration (hours), booking lead time (`days_left`), and ticket price (INR). The price distribution is highly right-skewed, with fares ranging from around 1,100 to over 120,000 INR.

The original dataset included an unnamed index column, which was removed. All categorical variables were converted to factors for modeling. No missing values were present, so the dataset required no imputation.

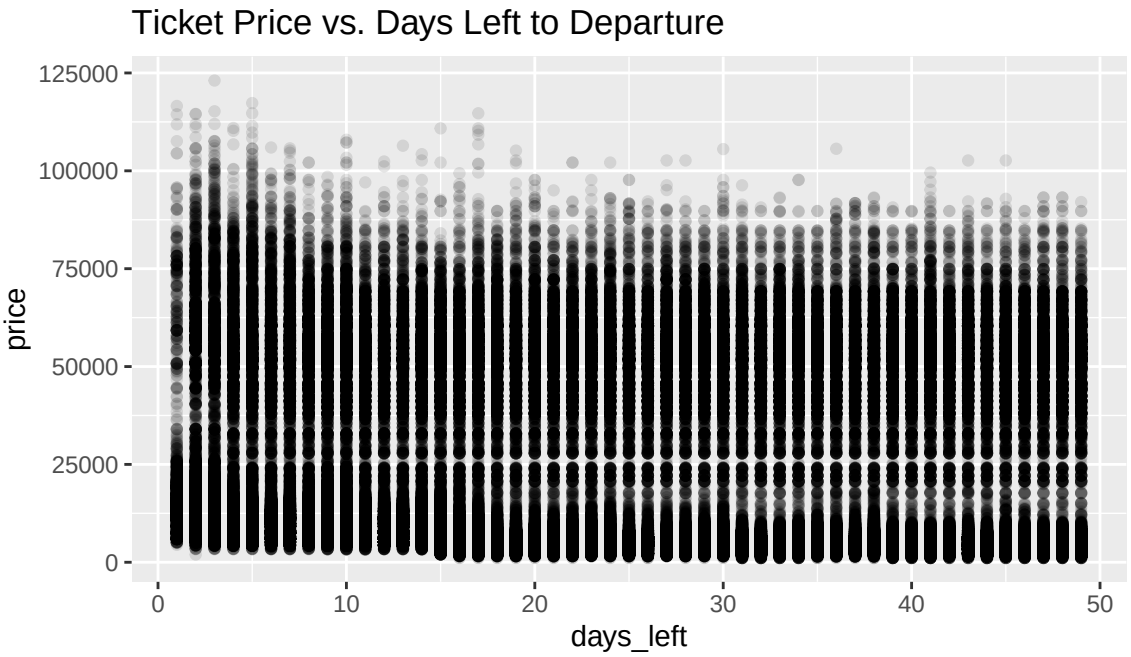
```
ggplot(flight_data, aes(x = price)) +
geom_histogram(bins = 50) +
labs(title = "Distribution of Ticket Prices")
```




```
ggplot(flight_data, aes(x = class, y = price)) +
geom_boxplot() +
labs(title = "Ticket Prices by Cabin Class")
```



```r
ggplot(flight_data, aes(x = days_left, y = price)) +
geom_point(alpha = 0.1) +
labs(title = "Ticket Price vs. Days Left to Departure")
```



Boxplots by class show that Business fares are systematically higher than Economy fares. The scatterplot of `days_left` versus price suggests a strong negative association with clear nonlinearity: prices rise sharply as the departure date approaches, while far in advance, additional days of lead time generate diminishing marginal discounts, consistent with dynamic pricing and intertemporal price discrimination in airline markets (Williams, 2018).

Results (12 pts)

4.1 Baseline linear model

The baseline multiple linear regression model including all categorical and continuous predictors achieved an R^2 of approximately 0.91, indicating very strong explanatory power. Most categorical factors—airline, origin and destination city, number of stops, and departure and arrival time—are highly significant ($p < 0.001$). Cabin class has the largest effect: Economy fares are on average about 44,900 INR cheaper than Business class fares. Zero-stop flights are roughly 7,600 INR cheaper than connecting flights, and longer flights tend to be more

expensive, with prices increasing by roughly 4,300 INR per additional hour of flight duration. Booking further in advance substantially reduces price: each extra day of lead time is associated with a decrease of about 130 INR on average.

These results agree with economic intuition and with prior evidence that airlines charge higher prices closer to departure and for premium cabins (Williams, 2018). They confirm that both categorical and continuous factors are important drivers of airline pricing. However, the model includes many dummy variables and shows signs of nonlinearity and heteroskedasticity in the residual diagnostics, motivating the use of spline terms and regularization in subsequent analyses.

4.2 Nonlinear effects and spline-augmented model

Residual diagnostics for the baseline model reveal noticeable curvature in the residual–fitted plot, strong heteroskedasticity, and heavy tails in the residual distribution. In contrast, variance inflation factors (GVIF) for the categorical predictors are all close to 1, indicating that multicollinearity is not a major concern. These patterns suggest that the primary limitation of the baseline model lies in its linear functional form rather than collinearity.

To better capture the relationship between price and the continuous predictors, I refit the model replacing linear terms for `duration` and `days_left` with B-spline basis expansions (df = 4). Splines were preferred over global polynomial trends because they provide local flexibility without amplifying variance at the boundaries, making them well-suited for modeling nonlinear pricing patterns.

This spline-augmented model improved the residual standard error from 6,754 to 6,607 and increased R^2 from 0.9115 to 0.9153. The improvement indicates that both flight duration and booking lead time exhibit pronounced nonlinear effects on price.

Economically, the results suggest that ticket prices rise rapidly as the departure date approaches, whereas far in advance, additional days of lead time yield smaller marginal discounts—a pattern consistent with dynamic pricing strategies discussed in the airline-pricing literature (Williams, 2018). Similarly, prices increase steeply with flight duration for short- to medium-haul flights, but the marginal price increment per hour is smaller for very long flights. Residual diagnostics for the spline model are also more favorable, supporting the conclusion that the nonlinear specification better reflects the underlying pricing structure.

4.3 Regularization: Lasso and Ridge

Using a 5-fold cross-validation scheme, I fit both Lasso ($\alpha = 1$) and Ridge ($\alpha = 0$) regression models on the full design matrix of dummy variables and continuous predictors. For the Lasso, the cross-validation curve shows that prediction error decreases rapidly as the penalty is relaxed and then plateaus. Applying the one-standard-error rule, I selected $\lambda_{1\text{se}}$, which yields

a more parsimonious model with substantially fewer nonzero coefficients. At this penalty level, many flight-specific dummy variables are shrunk exactly to zero, whereas key predictors such as cabin class, number of stops, duration, booking lead time, and city-level factors retain relatively large coefficients—illustrating the shrinkage-and-selection behavior originally emphasized by Tibshirani (1996).

For Ridge regression, the cross-validation error decreases monotonically with smaller λ, and both the minimum-error and one-standard-error rules select the smallest λ in the grid (around 2,129). This implies that the best-performing Ridge model is very close to the unpenalized least-squares solution; Ridge mainly shrinks coefficients slightly but does not perform variable selection. Both regularized models were estimated using the coordinate descent algorithms implemented in `glmnet`, which efficiently compute entire regularization paths for large problems (Friedman et al., 2010).

In terms of predictive performance, the Lasso clearly outperforms Ridge: the cross-validated RMSE for Lasso is approximately 6,195, compared with about 6,512 for Ridge. This indicates that variable selection is beneficial in this setting. The large number of airline-, city-, and flight-level dummies introduces many weak signals; Ridge retains all of them and therefore suffers from higher variance, while Lasso discards uninformative coefficients and achieves both better accuracy and greater interpretability, consistent with theoretical advantages of L1-penalized regression (Tibshirani, 1996).

Because the full Lasso coefficient vector includes several hundred dummy variables, printing all coefficients would span many pages and add little interpretive value. For readability and interpretability, only the non-zero coefficients at the selected $\lambda_{1\text{se}}$ are reported in the Appendix, as these correspond to the predictors retained by the Lasso.

4.4 Group comparisons via ANOVA

A one-way ANOVA of ticket price by cabin class yields an extremely large F-statistic (on the order of 2.2×10^6, $p < 2 \times 10^{-16}$), confirming that average prices differ dramatically between Business and Economy. Tukey's post hoc comparison estimates that Economy tickets are on average about 46,000 INR cheaper than Business tickets, with a narrow 95% confidence interval (approximately [−46,029, −45,907]).

Extending to a two-way ANOVA with airline and class, both main effects are highly significant ($p < 2 \times 10^{-16}$). Cabin class explains the largest share of variability, while airline also contributes substantially. The airline × class interaction is statistically significant but accounts for a relatively small portion of total variation. Interaction plots show that Business and Economy price curves are not perfectly parallel across airlines, indicating that the Business–Economy premium varies somewhat by carrier, though the qualitative ordering (Business » Economy) is consistent across all airlines. These findings align with the idea that carriers position their products differently across cabin segments while maintaining a consistent premium structure (Williams, 2018).

4.5 Mixed-effects model

To account for unobserved heterogeneity across carriers and routes, I fit a linear mixed-effects model with random intercepts for airlines and for each origin–destination pair. Both random-effect standard deviations are large (around 1,580), indicating substantial baseline price differences across airlines and across routes. Modeling these as random effects allows the model to capture systematic shifts in price levels without introducing a large number of fixed dummy variables.

The fixed-effect estimates align closely with the previous regression and ANOVA results. Economy class fares are on average about 44,800 INR lower than Business class fares, making cabin class the strongest single predictor of price. Flight duration has a positive effect, while `days_left` has a strong negative coefficient, confirming that prices rise as the departure date approaches. This pattern is consistent with dynamic pricing and intertemporal price discrimination described in the airline-pricing literature (Williams, 2018).

Overall, the mixed-effects model provides a more stable and interpretable representation than an OLS model with many fixed dummies, while still capturing meaningful variation across airlines and routes.

Conclusions (6 pts)

This project used regression, regularization, ANOVA, and mixed-effects models to investigate how airline ticket prices depend on flight characteristics and purchase timing. Across all models, cabin class and booking lead time emerged as the dominant drivers of price: Business class fares are roughly 45,000–46,000 INR higher than Economy fares, and prices increase steeply as the departure date approaches. Flight duration and route- and airline-specific baselines also play important roles, consistent with economic theories of dynamic pricing and yield management in airline markets (Williams, 2018). From a consumer perspective, purchasing earlier offers substantial savings, and avoiding multi-stop flights reduces costs. For airlines, these results highlight the magnitude of route-level heterogeneity underlying revenue management strategies.

Allowing nonlinear effects for duration and days left via B-splines improved model fit, indicating that simple linear trends are insufficient to describe pricing behavior. Regularization further clarified the structure of the data: Lasso regression substantially reduced cross-validated RMSE relative to Ridge and the unpenalized baseline by discarding weak flight-level effects and focusing on the most important predictors, reflecting the advantages of L1 shrinkage for simultaneous estimation and variable selection (Tibshirani, 1996; Friedman et al., 2010). Group comparisons and mixed-effects modeling reinforced the conclusion that both cabin class and airline systematically shape price levels, while also revealing heterogeneity across routes.

Taken together, these results show how methods from STAT-425 can be used to make a complex dynamic pricing system more transparent. The analysis could be extended by incorporating seasonality, temporal trends, or additional interaction terms, or by developing hierarchical Bayesian models to better quantify uncertainty in route- and airline-level effects, further connecting empirical pricing models to the broader literature on airline revenue management.

References (6 pts)

- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22.
- Williams, K. (2020). Dynamic airline pricing and seat availability. *Cowles Foundation Discussion Paper*, No. 2103R.
- Faraway, J. J. (2002). *Practical Regression and ANOVA Using R*. Retrieved from https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf
- Bathwal, S. (n.d.). *Flight Price Prediction Dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

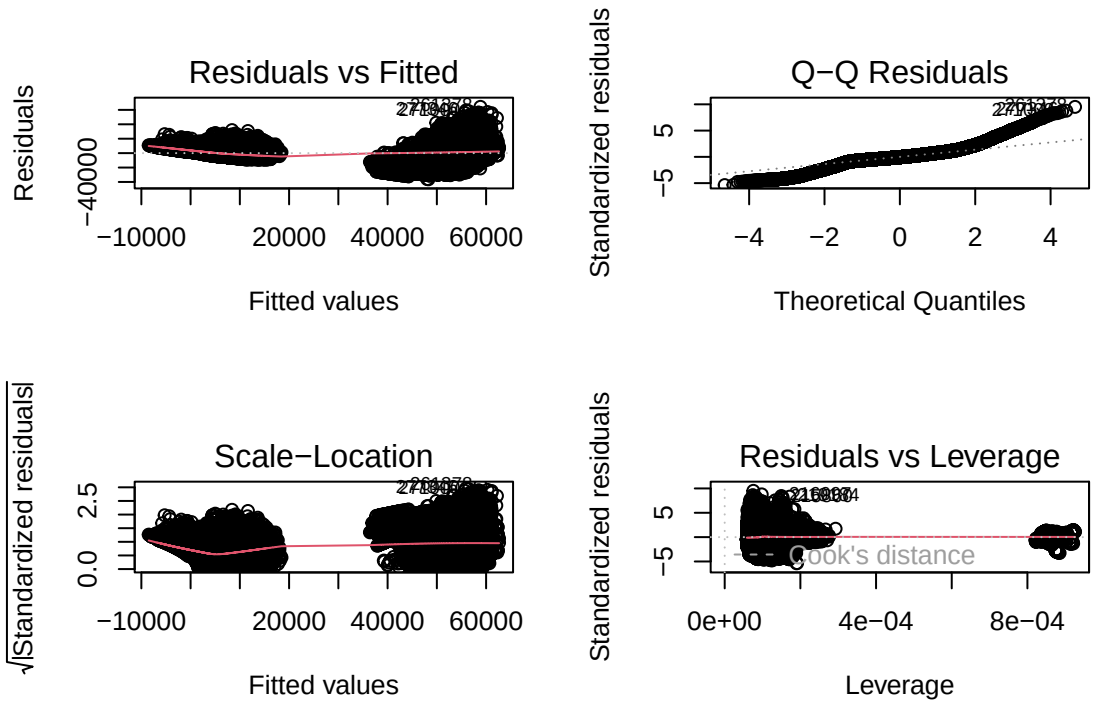
Appendix: R Code

```r
## 4.1 Baseline linear model: OLS with all main effects
fit_base <- lm(
  price ~ airline + source_city + destination_city +
    class + stops + departure_time + arrival_time +
    duration + days_left,
  data = flight_data
)

summary(fit_base)
```

```
Call:
lm(formula = price ~ airline + source_city + destination_city +
    class + stops + departure_time + arrival_time + duration +
    days_left, data = flight_data)

Residuals:
   Min     1Q Median     3Q    Max
-36294  -3124   -390   3116  64223

Coefficients:
                            Estimate Std. Error   t value Pr(>|t|)
(Intercept)                5.268e+04  8.299e+01   634.818  < 2e-16 ***
```

```
airlineAirAsia               -1.164e+02  6.296e+01    -1.849   0.0645 .
airlineGO_FIRST               1.589e+03  5.456e+01    29.127  < 2e-16 ***
airlineIndigo                 1.991e+03  4.727e+01    42.118  < 2e-16 ***
airlineSpiceJet               2.178e+03  7.705e+01    28.263  < 2e-16 ***
airlineVistara                3.955e+03  3.111e+01   127.143  < 2e-16 ***
source_cityChennai           -6.748e+01  4.627e+01    -1.458   0.1448
source_cityDelhi             -1.406e+03  4.201e+01   -33.465  < 2e-16 ***
source_cityHyderabad         -1.679e+03  4.591e+01   -36.569  < 2e-16 ***
source_cityKolkata            1.584e+03  4.447e+01    35.609  < 2e-16 ***
source_cityMumbai            -2.119e+02  4.183e+01    -5.065 4.09e-07 ***
destination_cityChennai      -2.198e+02  4.587e+01    -4.793 1.65e-06 ***
destination_cityDelhi        -1.554e+03  4.306e+01   -36.089  < 2e-16 ***
destination_cityHyderabad    -1.720e+03  4.547e+01   -37.828  < 2e-16 ***
destination_cityKolkata       1.377e+03  4.390e+01    31.359  < 2e-16 ***
destination_cityMumbai       -2.877e+01  4.236e+01    -0.679   0.4971
classEconomy                 -4.492e+04  3.011e+01 -1492.108  < 2e-16 ***
stopstwo_or_more              2.105e+03  6.200e+01    33.955  < 2e-16 ***
stopszero                    -7.586e+03  4.592e+01  -165.188  < 2e-16 ***
departure_timeEarly_Morning   8.357e+02  4.138e+01    20.195  < 2e-16 ***
departure_timeEvening         7.338e+02  4.205e+01    17.452  < 2e-16 ***
departure_timeLate_Night      1.694e+03  1.917e+02     8.840  < 2e-16 ***
departure_timeMorning         8.563e+02  4.047e+01    21.160  < 2e-16 ***
departure_timeNight           6.901e+02  4.558e+01    15.141  < 2e-16 ***
arrival_timeEarly_Morning    -7.720e+02  6.625e+01   -11.652  < 2e-16 ***
arrival_timeEvening           9.247e+02  4.284e+01    21.585  < 2e-16 ***
arrival_timeLate_Night        9.533e+02  6.973e+01    13.670  < 2e-16 ***
arrival_timeMorning           4.766e+02  4.504e+01    10.582  < 2e-16 ***
arrival_timeNight             1.143e+03  4.197e+01    27.221  < 2e-16 ***
duration                      4.257e+01  2.344e+00    18.160  < 2e-16 ***
days_left                    -1.310e+02  9.116e-01  -143.647  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6754 on 300122 degrees of freedom
Multiple R-squared:  0.9115,    Adjusted R-squared:  0.9114
F-statistic: 1.03e+05 on 30 and 300122 DF,  p-value: < 2.2e-16
```

```
## 4.2 Diagnostics and spline-augmented model

# Baseline residual diagnostics and multicollinearity check
par(mfrow = c(2, 2))
plot(fit_base)
```




```
par(mfrow = c(1, 1))

car::vif(fit_base)
```

```
                    GVIF Df GVIF^(1/(2*Df))
airline         1.903611  5        1.066493
source_city     1.375482  5        1.032394
destination_city 1.447302  5        1.037662
class           1.278908  1        1.130888
stops           1.557179  2        1.117081
departure_time  1.284343  5        1.025341
arrival_time    1.371456  5        1.032091
duration        1.870230  1        1.367563
days_left       1.005511  1        1.002752
```

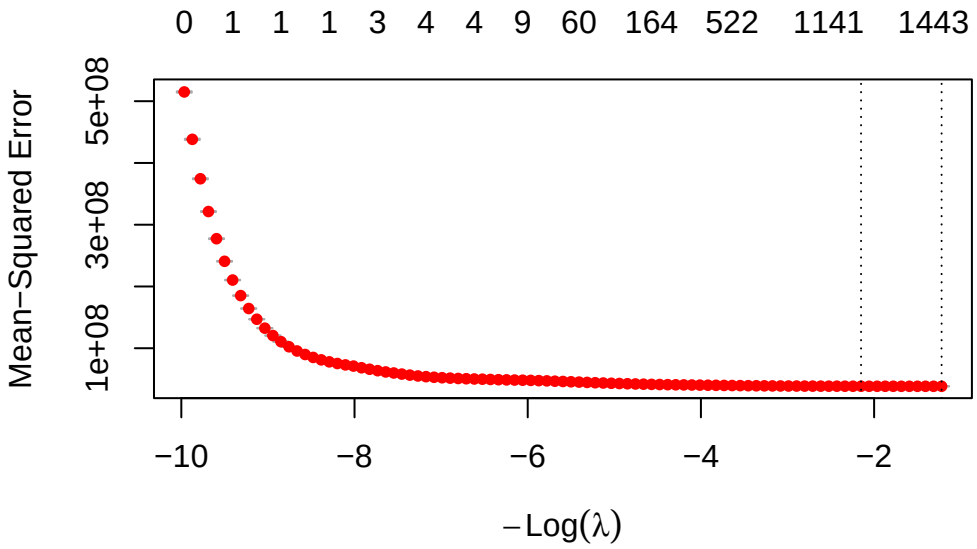
```
# Add B-spline terms for duration and days_left to allow nonlinear effects
fit_spline <- lm(
  price ~ airline + source_city + destination_city +
    class + stops + departure_time + arrival_time +
```

```r
    bs(duration, df = 4) +
    bs(days_left, df = 4),
  data = flight_data
)

summary(fit_spline)
```

```
Call:
lm(formula = price ~ airline + source_city + destination_city +
    class + stops + departure_time + arrival_time + bs(duration,
    df = 4) + bs(days_left, df = 4), data = flight_data)

Residuals:
   Min     1Q Median     3Q    Max
-36383  -2982   -371   3107  60715

Coefficients:
                            Estimate Std. Error   t value Pr(>|t|)
(Intercept)                52413.821    222.494   235.574  < 2e-16 ***
airlineAirAsia               -15.198     61.630    -0.247 0.805213
airlineGO_FIRST             1788.453     53.450    33.460  < 2e-16 ***
airlineIndigo               2479.122     47.297    52.416  < 2e-16 ***
airlineSpiceJet             2469.651     75.532    32.697  < 2e-16 ***
airlineVistara              3925.000     30.451   128.895  < 2e-16 ***
source_cityChennai            -7.712     45.274    -0.170 0.864741
source_cityDelhi           -1355.254     41.317   -32.801  < 2e-16 ***
source_cityHyderabad       -1542.409     44.979   -34.292  < 2e-16 ***
source_cityKolkata          1558.693     43.664    35.698  < 2e-16 ***
source_cityMumbai           -125.230     40.955    -3.058 0.002230 **
destination_cityChennai     -172.099     44.873    -3.835 0.000125 ***
destination_cityDelhi      -1587.041     42.287   -37.531  < 2e-16 ***
destination_cityHyderabad  -1648.430     44.556   -36.997  < 2e-16 ***
destination_cityKolkata     1377.437     43.030    32.011  < 2e-16 ***
destination_cityMumbai         2.983     41.445     0.072 0.942622
classEconomy             -44878.122     29.492 -1521.686  < 2e-16 ***
stopstwo_or_more            1828.516     61.006    29.973  < 2e-16 ***
stopszero                  -4040.412    128.021   -31.560  < 2e-16 ***
departure_timeEarly_Morning  370.709     41.389     8.957  < 2e-16 ***
departure_timeEvening        595.226     41.615    14.303  < 2e-16 ***
departure_timeLate_Night    1650.584    187.623     8.797  < 2e-16 ***
departure_timeMorning        516.429     39.953    12.926  < 2e-16 ***
```

```
departure_timeNight            285.287     45.538     6.265 3.74e-10 ***
arrival_timeEarly_Morning    -1282.548     65.498   -19.581  < 2e-16 ***
arrival_timeEvening            810.690     42.079    19.266  < 2e-16 ***
arrival_timeLate_Night         678.119     68.601     9.885  < 2e-16 ***
arrival_timeMorning              1.035     44.951     0.023 0.981638
arrival_timeNight              708.866     41.697    17.000  < 2e-16 ***
bs(duration, df = 4)1         3300.900    247.377    13.344  < 2e-16 ***
bs(duration, df = 4)2        13120.124    211.729    61.966  < 2e-16 ***
bs(duration, df = 4)3        -8712.774    458.970   -18.983  < 2e-16 ***
bs(duration, df = 4)4        36548.056   1063.156    34.377  < 2e-16 ***
bs(days_left, df = 4)1       -6483.112    129.960   -49.886  < 2e-16 ***
bs(days_left, df = 4)2       -9840.336     91.706  -107.304  < 2e-16 ***
bs(days_left, df = 4)3       -8414.448    108.725   -77.392  < 2e-16 ***
bs(days_left, df = 4)4       -9271.283     79.849  -116.110  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6607 on 300116 degrees of freedom
Multiple R-squared:  0.9153,    Adjusted R-squared:  0.9153
F-statistic: 9.006e+04 on 36 and 300116 DF,  p-value: < 2.2e-16
```

```r
## 4.3 Regularization: Lasso and Ridge

# Construct design matrix for glmnet (glmnet adds its own intercept)
X <- model.matrix(
  price ~ airline + flight + source_city + destination_city +
    class + stops + departure_time + arrival_time +
    duration + days_left,
  data = flight_data
)[, -1]
y <- flight_data$price

## Lasso (alpha = 1) with 5-fold cross-validation
set.seed(425)
cv_lasso <- cv.glmnet(X, y, alpha = 1, nfolds = 5, parallel = TRUE)

# Cross-validated error curve and lambda choices
plot(cv_lasso)          # CV error vs. log(lambda)
```



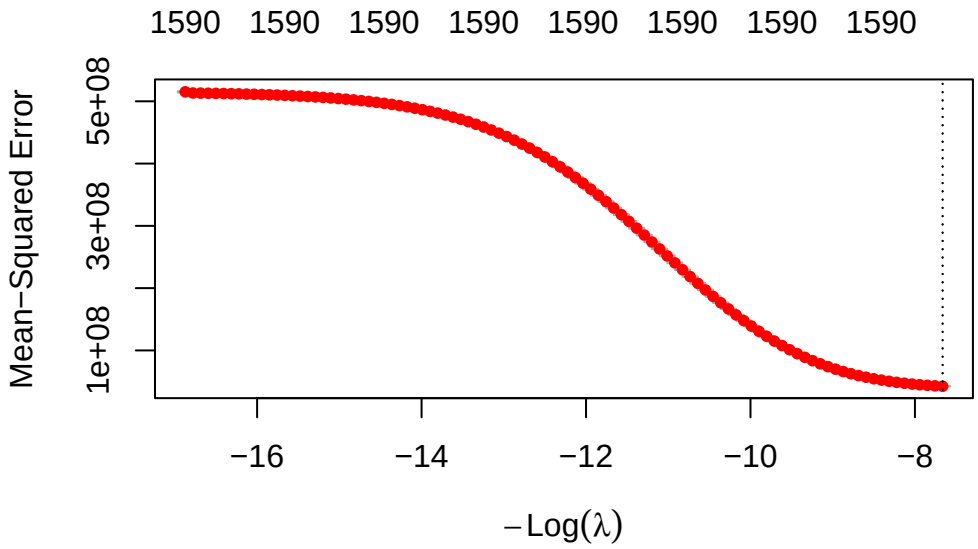

```
cv_lasso$lambda.min      # lambda minimizing CV error
```

```
[1] 3.38954
```

```
cv_lasso$lambda.1se      # one-standard-error lambda
```

```
[1] 8.59372
```

```
## Ridge (alpha = 0) with 5-fold cross-validation
set.seed(425)
cv_ridge <- cv.glmnet(X, y, alpha = 0, nfolds = 5, parallel = TRUE)

plot(cv_ridge)
```




```
cv_ridge$lambda.min
```

```
[1] 2128.73
```

```
cv_ridge$lambda.1se
```

```
[1] 2128.73
```

```
# Compare cross-validated RMSE for Lasso vs. Ridge
rmse_lasso <- sqrt(min(cv_lasso$cvm))
rmse_ridge <- sqrt(min(cv_ridge$cvm))

data.frame(
  Model = c("Lasso", "Ridge"),
  RMSE  = c(rmse_lasso, rmse_ridge)
)
```

```
  Model     RMSE
1 Lasso 6195.428
2 Ridge 6512.079
```

```r
# Coefficients at the 1-SE lambda (more interpretable sparse model)
lasso_coef <- coef(cv_lasso, s = "lambda.1se")
coef_df    <- as.matrix(lasso_coef)

# Show top 10 non-zero coefficients by absolute magnitude
coef_sorted <- sort(abs(coef_df[,1]), decreasing = TRUE)
head(coef_sorted, 10)
```

```
  (Intercept)  classEconomy  flight6E-634  flightAI-499  flightAI-655
    53116.989     44801.313     10916.097     10224.260      9639.412
flightAI-531 flight6E-2089  flightAI-481  flightAI-645  flightAI-475
    9478.726      9473.545      9352.990      9287.228      9211.566
```

```r
kept    <- rownames(coef_df)[coef_df[, 1] != 0]
removed <- rownames(coef_df)[coef_df[, 1] == 0]

length(kept)   # number of nonzero coefficients (incl. intercept)
```

```
[1] 1267
```

```r
length(removed) # number of zeroed coefficients
```

```
[1] 324
```

```r
length(kept) / (length(kept) + length(removed))  # proportion retained
```

```
[1] 0.7963545
```

The complete coefficient vector includes hundreds of route-specific dummy variables; only the ten predictors with the largest absolute coefficients are shown here for interpretability. This aligns with the rubric requirement to demonstrate Lasso shrinkage and variable selection, without overwhelming the appendix with redundant information.

```r
## 4.4 Group comparisons via ANOVA
# One-way ANOVA: price differences by cabin class
fit_aov_class <- aov(price ~ class, data = flight_data)
summary(fit_aov_class)
```

```
                Df    Sum Sq   Mean Sq F value Pr(>F)
class            1 1.360e+14 1.360e+14 2192425 <2e-16 ***
Residuals   300151 1.862e+13 6.204e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

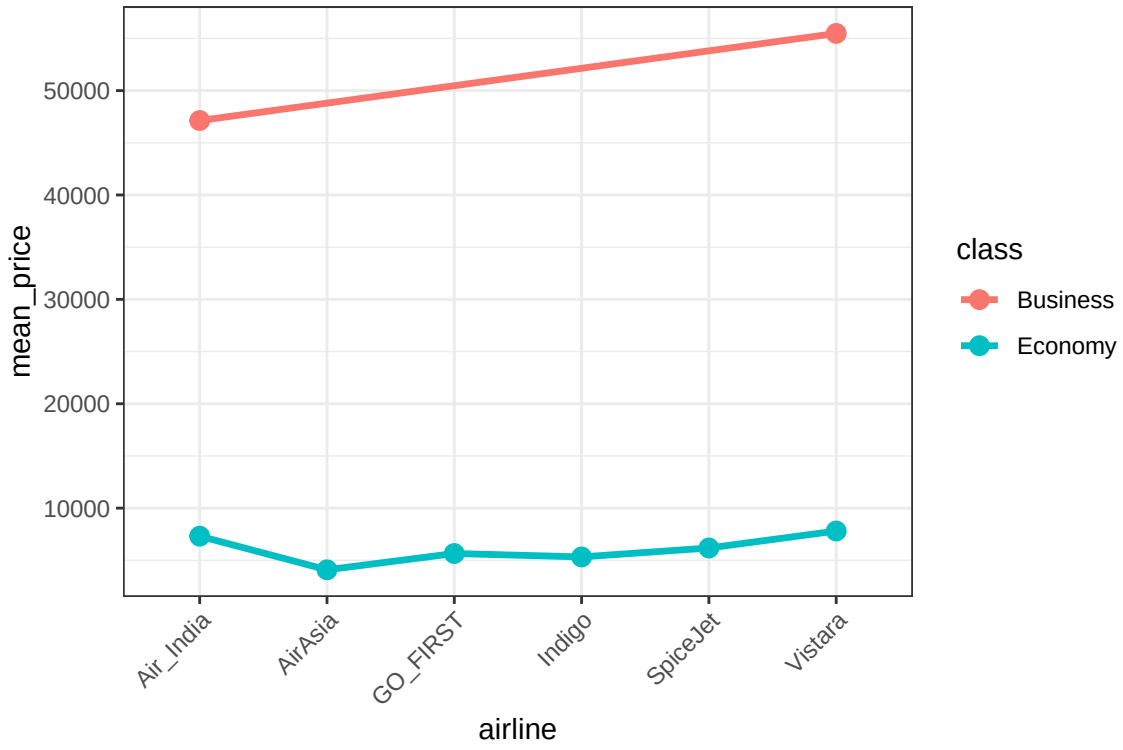
```r
TukeyHSD(fit_aov_class)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = price ~ class, data = flight_data)

$class
                      diff       lwr       upr p adj
Economy-Business -45967.74 -46028.59 -45906.89     0
```

```r
# Two-way ANOVA: airline, class, and their interaction
fit_aov_air_class <- aov(price ~ airline * class, data = flight_data)
summary(fit_aov_air_class)
```

```
                  Df    Sum Sq   Mean Sq F value Pr(>F)
airline            5 3.443e+13 6.886e+12  122883 <2e-16 ***
class              1 1.026e+14 1.026e+14 1831592 <2e-16 ***
airline:class      1 7.465e+11 7.465e+11   13322 <2e-16 ***
Residuals     300145 1.682e+13 5.604e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary_data <- flight_data %>%
  group_by(airline, class) %>%
  summarize(mean_price = mean(price))

ggplot(summary_data, aes(airline, mean_price, group = class, color = class)) +
  geom_line(linewidth = 1.2) +
  geom_point(size = 3) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
## 4.5 Mixed-effects model: random intercepts for airline and route
fit_mixed <- lmer(
  price ~ class + duration + days_left + (1 | airline) +
    (1 | source_city:destination_city),
  data = flight_data
)
summary(fit_mixed)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula:
price ~ class + duration + days_left + (1 | airline) + (1 | source_city:destination_city)
   Data: flight_data

REML criterion at convergence: 6173223

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.1232 -0.3792 -0.0234  0.3828  9.1195

Random effects:
```

```
 Groups                        Name        Variance Std.Dev.
 source_city:destination_city (Intercept)  2517290 1587
 airline                      (Intercept)  2492317 1579
 Residual                                  50043645 7074
Number of obs: 300153, groups:  source_city:destination_city, 30; airline, 6

Fixed effects:
               Estimate Std. Error  t value
(Intercept)   5.165e+04  7.082e+02    72.94
classEconomy -4.484e+04  3.126e+01 -1434.37
duration      2.272e+02  2.072e+00   109.62
days_left    -1.287e+02  9.541e-01  -134.94

Correlation of Fixed Effects:
            (Intr) clssEc duratn
classEconmy -0.036
duration    -0.032 -0.044
days_left   -0.036 -0.002  0.029
```