Econ 104 Project 3

Meghna Nair Ishika Agrawal Rajasvi Singh Anshika Khandelwal

2025-05-22

Question 1: Introduction of Data set

a)

This project uses a balanced panel data set containing annual observations for six major U.S. airlines over the period 1970 to 1984, resulting in a total of 90 observations. The data set captures several key economic and operational indicators for each airline across time. The aim of this study is to analyze how variations in output levels, fuel prices, and capacity utilization affect the total operating costs of these airlines. By modeling cost behavior using panel data techniques, the project seeks to uncover whether airlines benefit from economies of scale and how external factors like fuel price fluctuations influence cost structures.

The key variables used in the analysis are:

1. cost: total operating cost (in thousands of dollars)
2. output: revenue passenger miles (index)
3. price: fuel price
4. load: average capacity utilization
5. firm: airline identifier
6. year: year of observation

The central research question guiding this project is: How do output, fuel price, and load factor influence the operating costs of U.S. airlines? The findings are expected to shed light on cost-efficiency strategies within the airline industry and provide evidence that could support data-driven policy or operational decisions.

Dataset Source: https://www.kaggle.com/datasets/sandhyakrishnan02/paneldata

```
##   I T       C        Q     PF       LF
## 1 1 1 1140640 0.952757 106650 0.534487
## 2 1 2 1215690 0.986757 110307 0.532328
## 3 1 3 1309570 1.091980 110574 0.547736
## 4 1 4 1511530 1.175780 121974 0.540846
## 5 1 5 1676730 1.160170 196606 0.591167
## 6 1 6 1823740 1.173760 265609 0.575417
```

b)

Summaries

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   68978  292046  637001 1122524 1345968 4748320
```

The values range from about 69K to 4.7M, with a median of 637K and a much higher average of 1.1M, showing strong right skew. Most airlines fall between 292K and 1.35M, but a few very large values are pulling the mean up.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03768 0.14213 0.30503 0.54499 0.94528 1.93646
```

This distribution is highly right-skewed, with a mean significantly greater than the median. Most values are concentrated at the lower end, indicating that many observations have low performance or output, while a few reach very high levels. The spread from 1st to 3rd quartile is wide, reinforcing variability across units.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  103795  129848  357434  471683  849840 1015610
```

Costs vary widely, with a long right tail (high maximum relative to mean and median). The distribution is right-skewed, and the large gap between median and 3rd quartile indicates that a subset of airlines incur disproportionately high costs.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4321  0.5288  0.5661  0.5605  0.5947  0.6763
```

This distribution is relatively symmetric and tightly clustered, with a small range between quartiles. The closeness of mean and median suggests no skewness. Overall, fuel price appears to be a stable variable across observations.

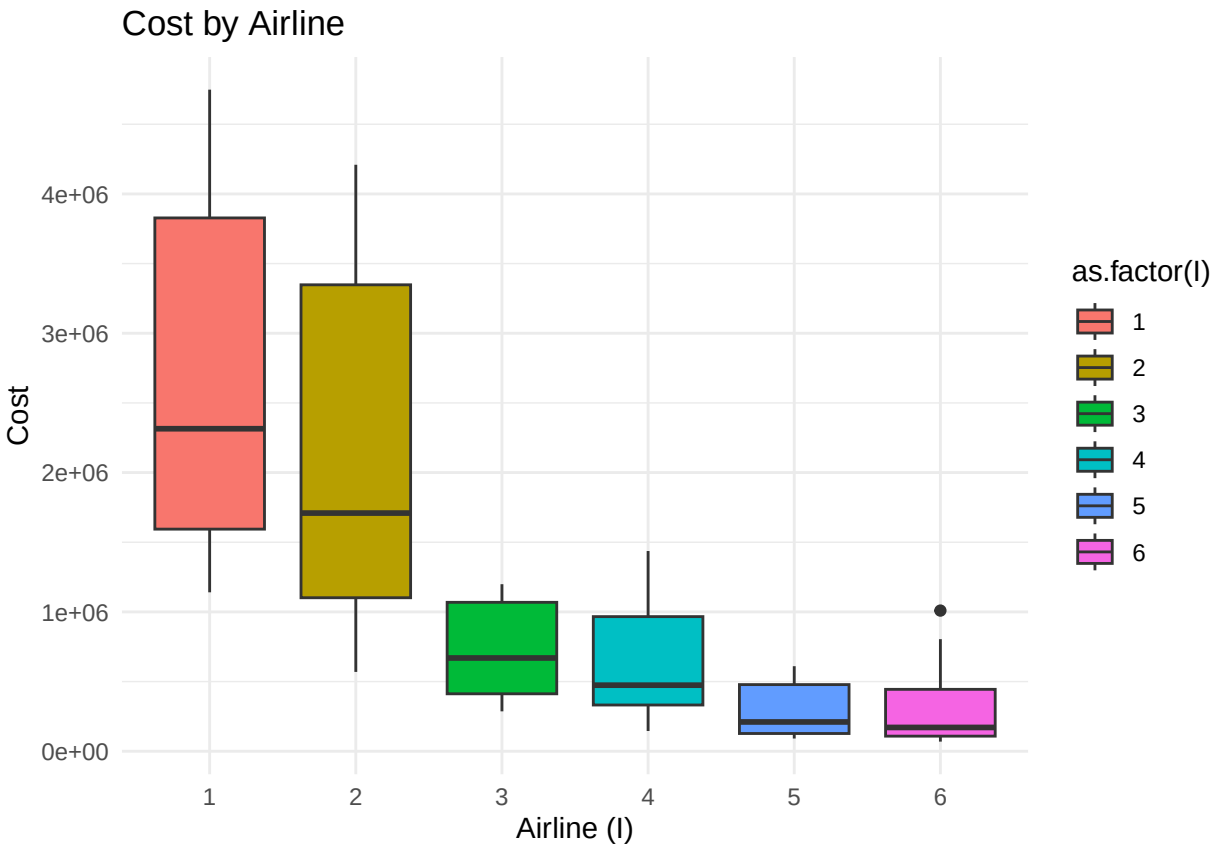
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     2.0     3.5     3.5     5.0     6.0
```

This represents a categorical variable treated numerically (possibly airline identifiers from 1 to 6). The uniform spacing of quartiles suggests a fairly balanced representation of different airlines in the dataset.

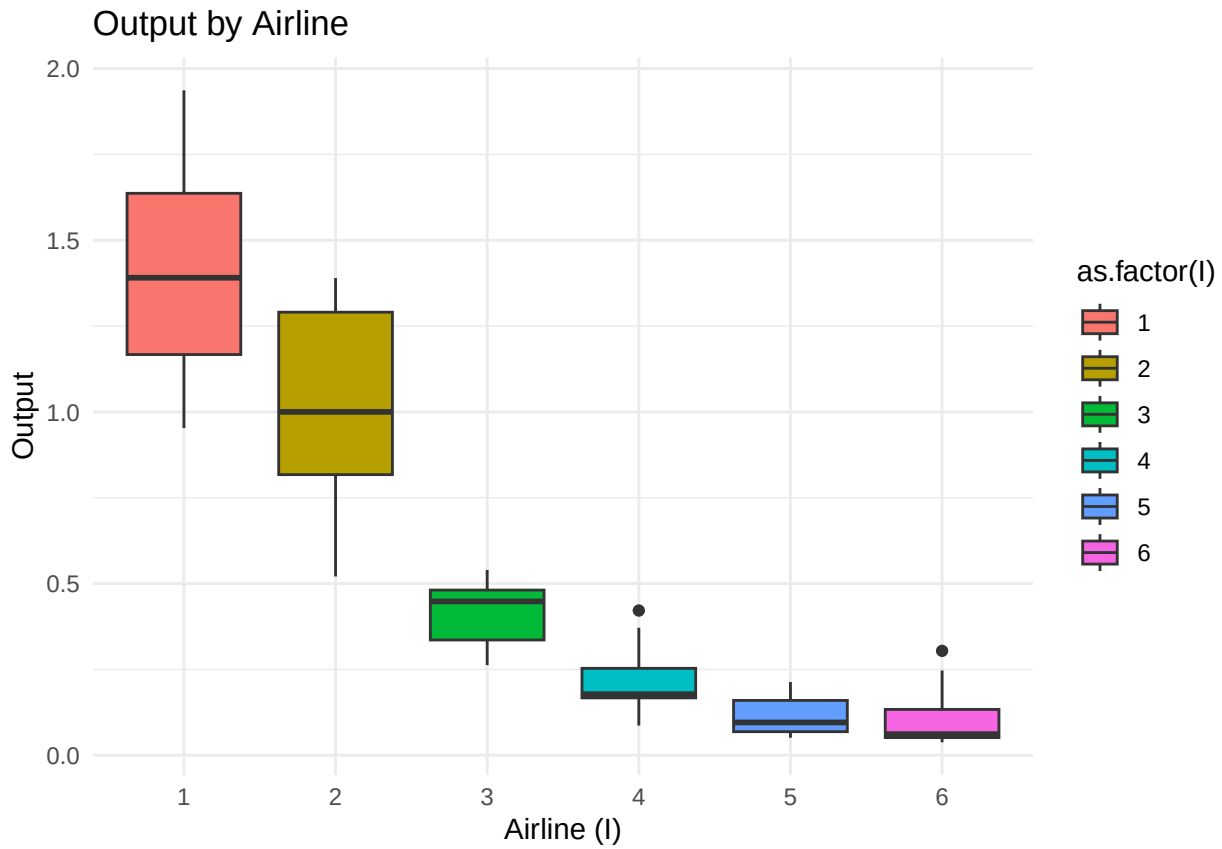
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       4       8       8      12      15
```

This discrete variable has a symmetric distribution. The mean equals the median, indicating balance and no skewness. The spread is moderate, indicating that most airlines are operating a similar number of flights or routes.

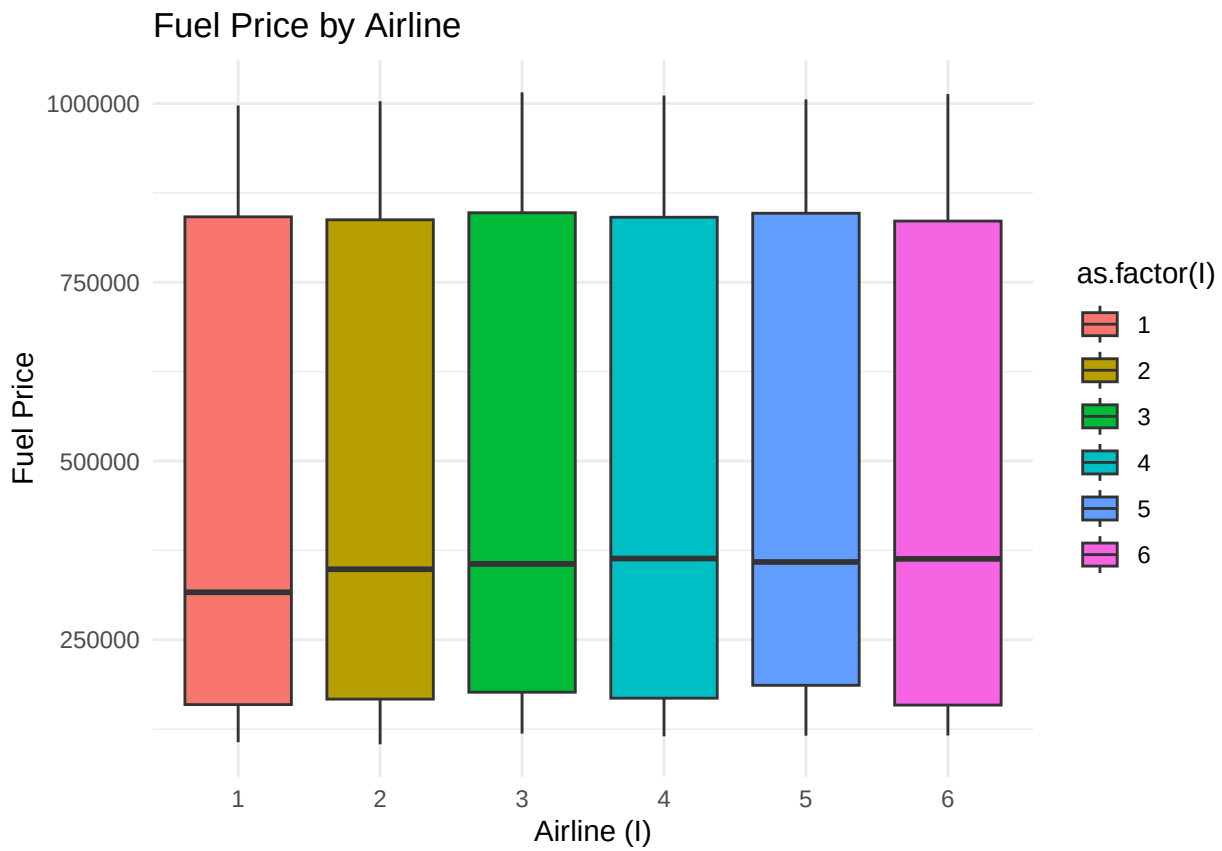
Box Plots



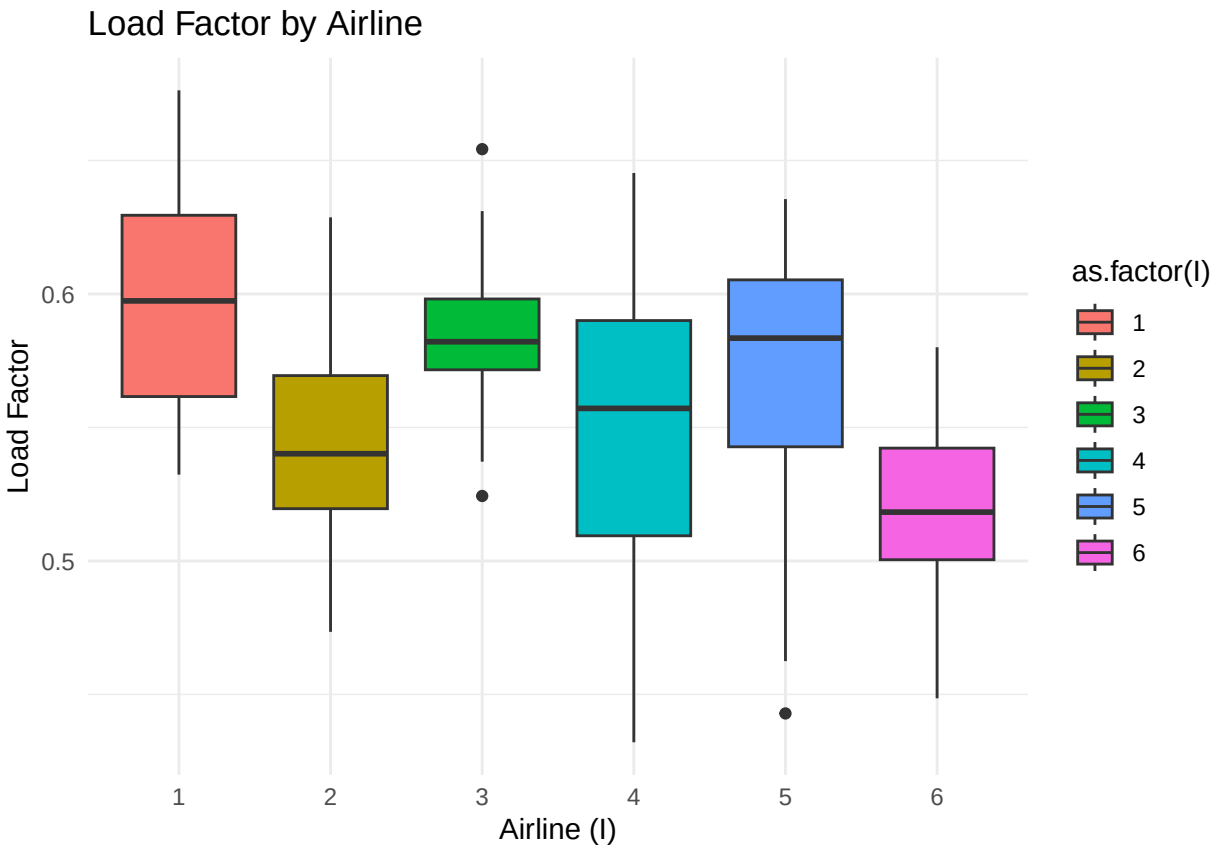
This boxplot reveals considerable heterogeneity in cost across airlines. Airlines 1 and 2 exhibit the highest median costs, with wide interquartile ranges (IQR), indicating large cost variability within these groups. This could be due to differences in route structures or scale of operations. Airlines 5 and 6 show significantly lower and more consistent costs, as suggested by their narrower IQRs. The distribution for most airlines appears right-skewed, especially for Airline 1, as indicated by the longer upper whiskers and presence of extreme values. This suggests some airlines may experience sporadically high cost spikes.



Output varies considerably across airlines, again suggesting individual heterogeneity. Airlines 1 and 2 not only produce more output on average (as seen from their higher medians) but also show a broader spread in output values. Airlines 3 through 6 have much lower median output and narrower IQRs, with some outliers, especially for Airlines 4 and 6. This suggests consistent low output among these airlines with occasional performance deviations. Like the cost data, the output distribution appears slightly right-skewed for several airlines.

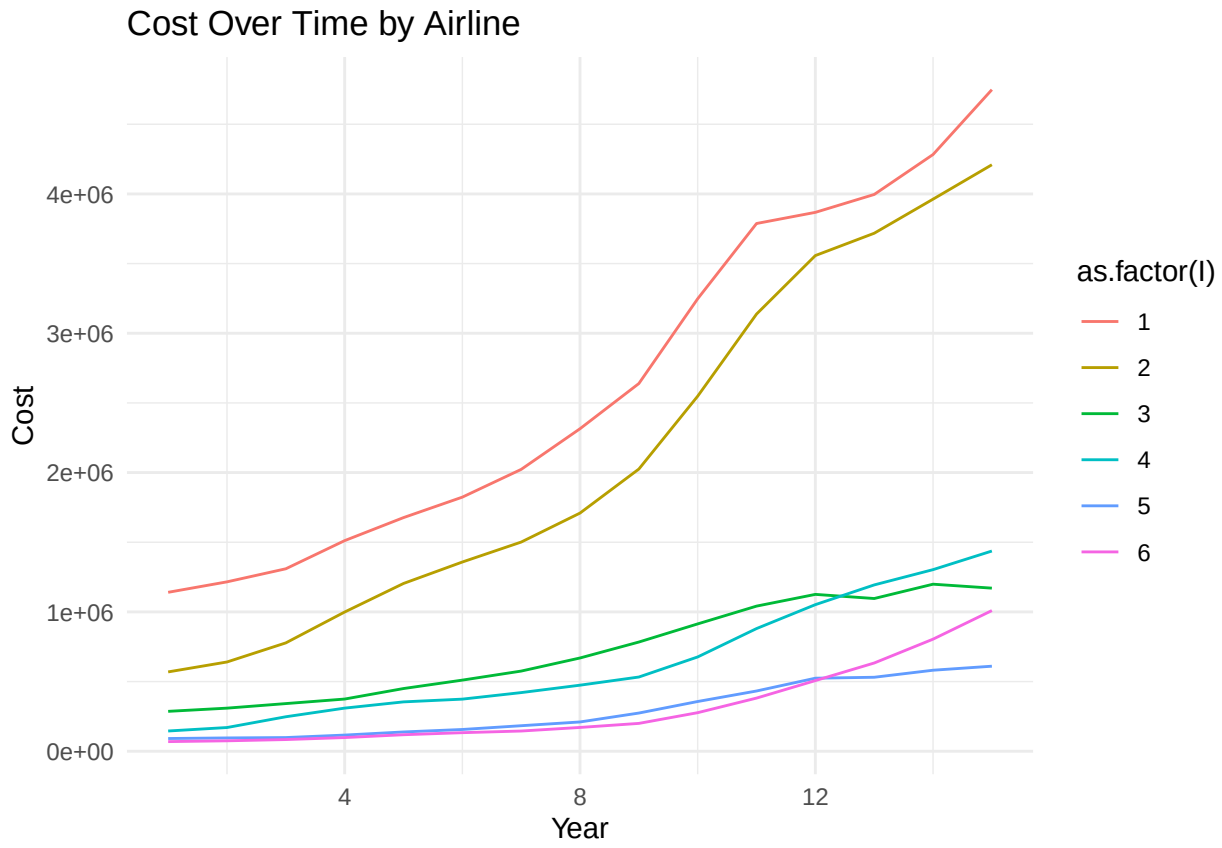


Fuel price distributions are remarkably consistent across all airlines, as shown by nearly identical boxplot shapes and medians. The symmetric and uniform spread suggests that fuel price is not subject to individual heterogeneity across airlines. This likely reflects that fuel prices are determined by external market forces and are not within the control of individual carriers. The distributions are roughly symmetric, with no significant skewness or outliers, indicating a stable variable over time and across units.

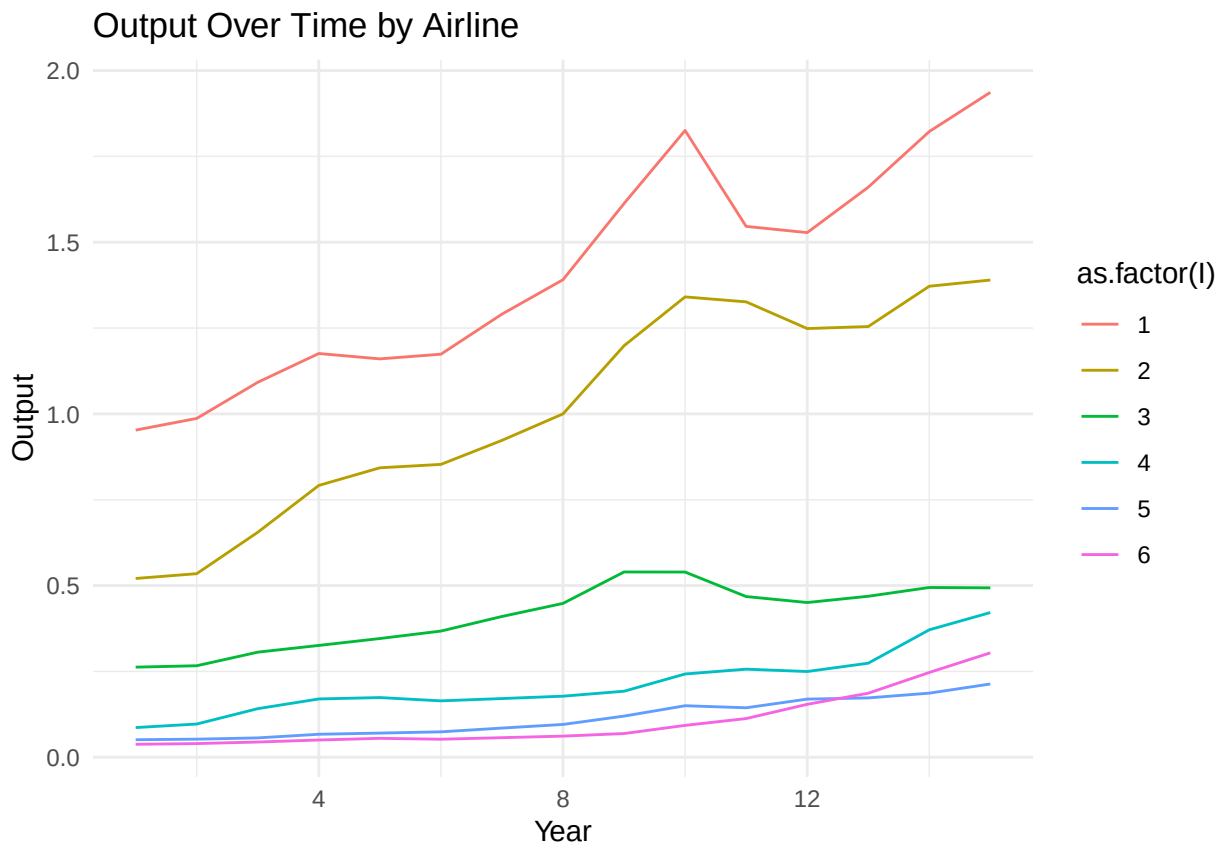


The boxplot displays the distribution of load factor across six different airlines (coded 1–6), providing insight into both central tendency and variability of performance. Airline 5 shows the highest median load factor (just above 0.60), indicating efficient seat utilization, while Airline 3, despite a slightly lower median, exhibits the most consistent performance with a very narrow interquartile range (IQR), suggesting little variation in load factor across observations. In contrast, Airline 4 has a wide IQR and long whiskers, implying substantial within-group variability or potential time heterogeneity, while Airlines 2 and 6 have the lowest median values (near 0.53–0.54), with Airline 2 also showing notable spread. Skew appears minimal for most airlines, though some mild asymmetry and outliers are present (especially in Airlines 3 and 5), which could reflect occasional under- or overperformance. Since this is a pooled cross-sectional boxplot (not time-separated), further plots like time series or scatterplots by airline would help confirm if the heterogeneity is due to differences across airlines or fluctuations within them over time.

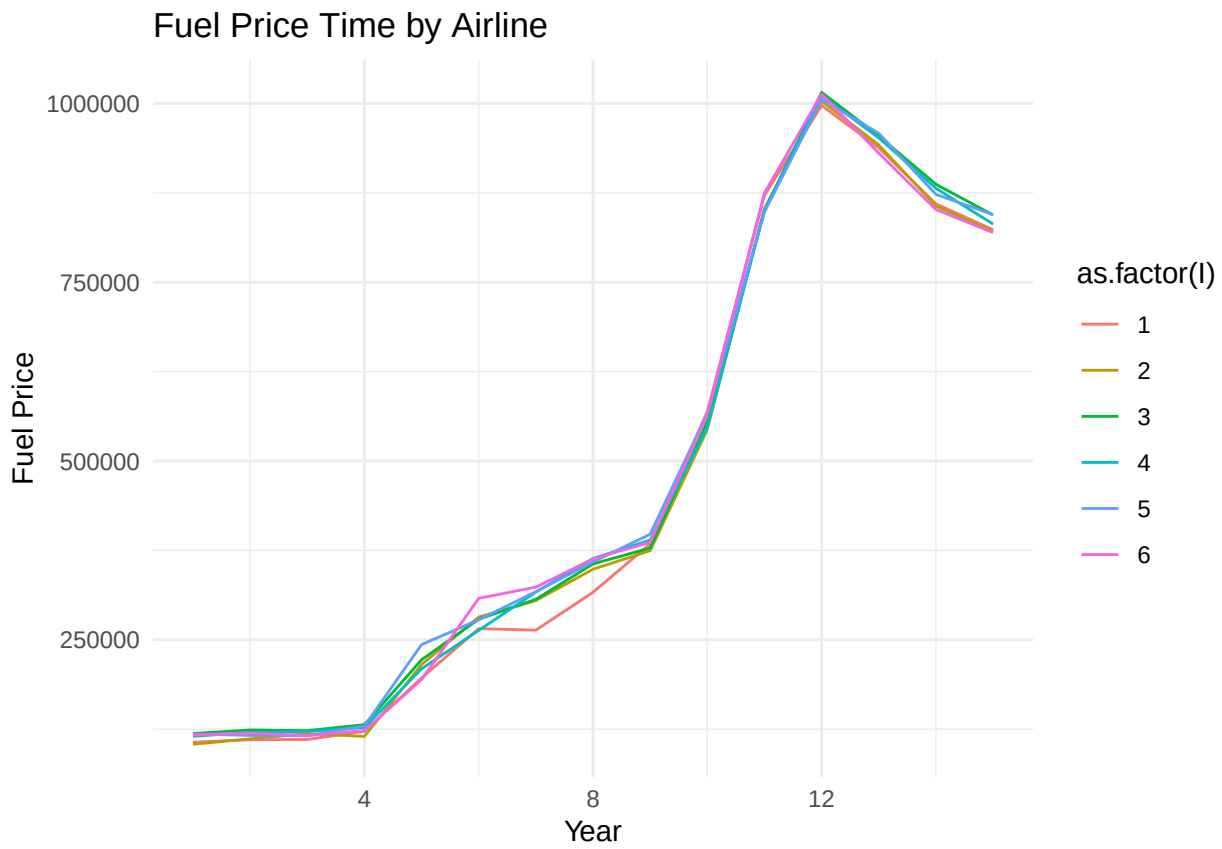
Line Plots



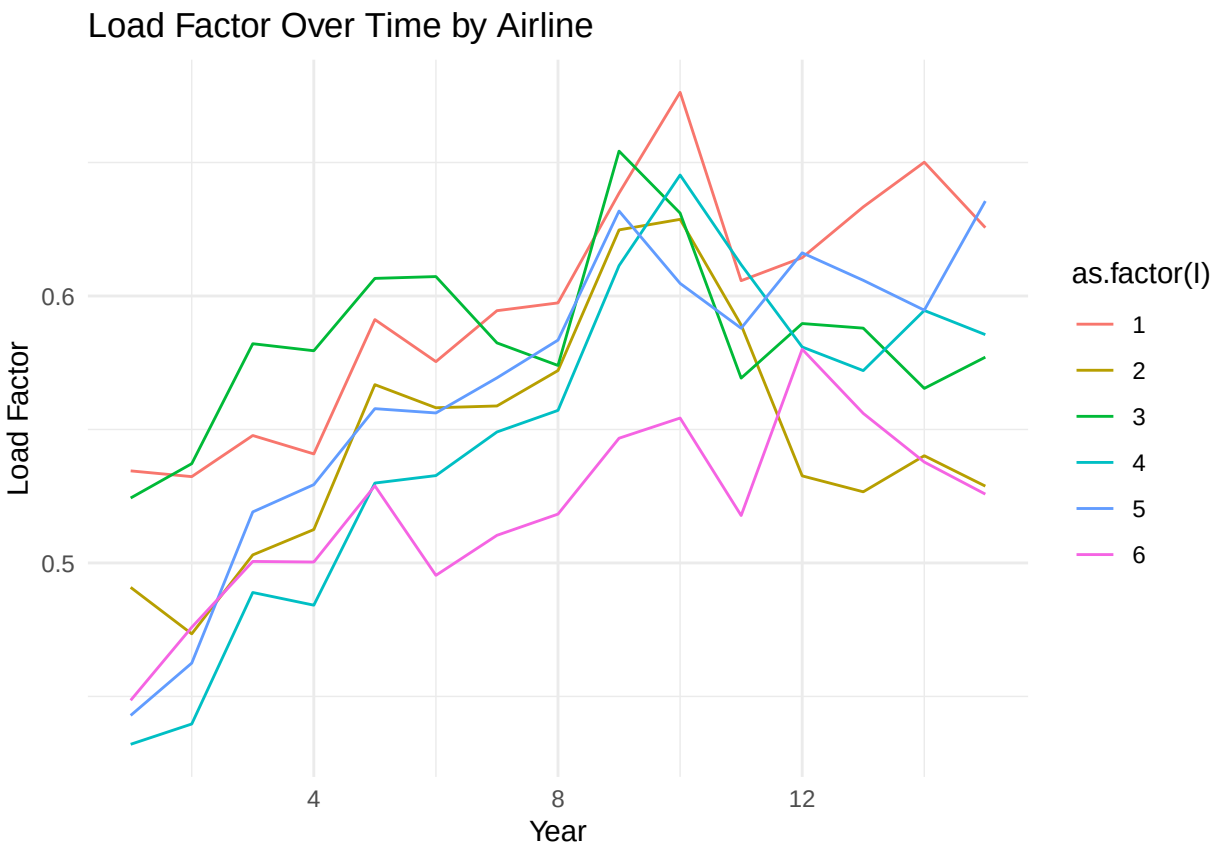
The line plot "Cost Over Time by Airline" shows that costs increase steadily over time for all six airlines, indicating clear time heterogeneity. Airlines 1 and 2 have the highest and fastest-growing costs, with Airline 1 reaching over 4 million units by the final year. Airlines 3 through 6 operate at much lower cost levels, with Airlines 5 and 6 showing the slowest growth. The crossover of Airlines 3 and 4 around year 11 suggests a shift in cost structure. Overall, the plot highlights both time and individual heterogeneity in costs, suggesting the need for models that account for differences across airlines and over time.



The plot shows output trends over time for six different airlines. Airline 1 consistently outperforms the others, with a steady increase in output and the highest overall values, peaking around year 15. Airline 2 also demonstrates strong growth, although it remains below airline 1 throughout the period. Airlines 3, 4, 5, and 6 exhibit much lower outputs, with relatively modest growth over time. The gap between the top two airlines and the rest suggests a significant performance disparity, possibly due to differences in scale, resources, or operational efficiency. Overall, the industry appears to show gradual growth, but it's concentrated in a few leading players.



The figure, "Fuel Price Over Time by Airline," reveals a consistent trend across all airlines, showing a sharp rise in fuel prices peaking around year 12, followed by a gradual decline. The close alignment of the lines for all six airlines suggests that fuel prices were largely uniform across the industry, likely driven by external market forces rather than airline-specific factors. This uniformity implies a shared sensitivity to global fuel market trends or similar procurement practices across the carriers.



The figure, "Load Factor Over Time by Airline," illustrates more variability among airlines. Load factors generally increased for all airlines until around year 10, after which they became more erratic. Airline 1 maintained a relatively high and stable load factor following the peak, consistent with its leading output performance observed in the first chart. In contrast, Airlines 2 and 6 experienced noticeable declines in load factor post-year 10, potentially signaling challenges in demand forecasting or operational efficiency. The divergence in load factor trends after the fuel price peak suggests that airlines responded differently to rising costs, possibly through changes in capacity or pricing strategies.

c) Model Fitting

Step 1: We checked if data is balanced using is.palanced function to ensure that our panel data set is appropriately balanced

Pooled Model

```
## [1] TRUE
```

We fit the pooled model, treating all observations as if they come from a single homogeneous group.

Fixed Effects Models

We fit the fixed effects model to control for unobserved heterogeneity across individuals (or entities) and over time. This allows us to isolate the impact of the explanatory variables on the dependent variable, while accounting for individual-specific or time-specific effects that might bias the results if ignored.

Comparison Between Pooled & Fixed Effects Models: Joint F-Test

```
##
##  F test for individual effects
##
## data:  C ~ Q + PF + LF
## F = 14.595, df1 = 5, df2 = 81, p-value = 3.467e-10
## alternative hypothesis: significant effects
```

```
##
##  F test for time effects
##
## data:  C ~ Q + PF + LF
## F = 0.56214, df1 = 14, df2 = 72, p-value = 0.8854
## alternative hypothesis: significant effects
```

The pF-test comparing the pooled model and the fixed effects model with individual-specific effects shows a highly significant result ($F = 14.60$, $p < 0.001$), indicating that there are important differences across the airline firms that need to be accounted for in the model. However, the pF-test comparing the pooled model and the fixed effects model with time-specific effects is not significant ($F = 0.56$, $p = 0.8854$), suggesting that variation over the years does not have a meaningful impact on the dependent variable once individual effects are considered. Therefore, the pooled OLS model and the fixed effect model with time- specific effects are inadequate, and the fixed effects model with individual-specific effects is more appropriate for our analysis.

Random Effects Model

Random Effects Test

```
##
##  Lagrange Multiplier Test - (Honda)
##
## data:  C ~ Q + PF + LF
## normal = 0.783, p-value = 0.2168
## alternative hypothesis: significant effects
```

The Lagrange Multiplier test for Random Effects yielded a test statistic of 0.783 with a p-value of 0.2168, indicating that we fail to reject the null hypothesis of no individual random effects. This suggests that there is no significant panel-level variance that would justify using a random effects model.

Hausman Test

```
##
##  Hausman Test
##
## data:  C ~ Q + PF + LF
## chisq = 60.87, df = 3, p-value = 3.832e-13
## alternative hypothesis: one model is inconsistent
```

The Hausman test produced a chi-squared statistic of 60.87 and a highly significant p-value of 3.832 x 10ˆ-13. This indicates strong evidence to reject the null hypothesis that the random effects model is consistent. Therefore, the fixed effects model is preferred over the random effects model for this data, as it provides consistent and reliable estimates.

d) Real World Implications

Our analysis shows that differences between individual airlines have a significant impact on the response variable (C), while changes over time don't matter as much. This means that factors unique to each firm (i.e. management quality, operational practices, market position, etc.) are key drivers of the results, rather than general trends across time. For policymakers, this suggests that focusing on improving individual performance (such as better management or higher operational standards) could be more effective than broad, time-based policies. For investors and stakeholders, understanding these individual-specific differences can help them make smarter decisions about where to allocate resources or how to value firms, by identifying those that consistently perform well regardless of the overall market. Using the fixed effects models with individual-specific effects to quantify these impacts helps us better understand the role of each airline's characteristics, which can translate into greater earning potential and competitive advantage. However, since the fixed effects model removes time-invariant variables, we should also consider implementing the Hausman-Taylor estimator to recover the lost effects of the time-invariant variables. This would allow us to estimate the influence of both individual-specific effects and of variables that do not vary over time, helping policymakers and investors assess the impact of stable firm characteristics alongside operational performance.