Individual Project

Data Visualization STAE04, Lund 22HT

Yiying(Linda) Ren

2022-10-27

Introduction

Data set

We will be working on a multivariate dataset called **Airline** from package `Ecdat` for this project for course Data Visualization. This is a dataset that contains 90 observations of a panel of 6 different airlines from 1970 to 1984 (duration of 15 years) with 6 variables in the United States.

References Greene, W.H. (2003) Econometric Analysis, Prentice Hall, https://archive.org/details/econometricanaly0000gree_f4x3, Table F7.1.

Table 1: Description of the variables in dataset Airline from package `Ecdat` (collected 1970-1984 in USA).

Variable	Data Type	Class	Description
airline	numerical	integer	id number of airline company
year	numerical	integer	the year data was collected
cost	numerical	integer	total cost, in $1,000
output	numerical	double	in revenue passenger miles, index number
pf	numerical	integer	fuel price
lf	numerical	double	load factor, the average capacity utilization of the fleet

Data Wrangling

This is a very clean and neat dataset without any missing values so that there is no need for data cleaning. However we will convert variable *airline* into a factor with 6 levels so that we can group data with later on in our analysis and also rename it to *ID* to differentiate from the name of the dataset. Also we will convert *year* from integer ranging 1 - 15 back to its original form as years 1970 - 1984 to be more informative. Lastly we will rename the variables *pf* to *fuel_price* and *lf* to *load* to

be more informative. After the data preparation, the first 6 observations of the new dataset *airline* are presented in the table below.

Table 2: Table of first 6 obserations after transformation

ID	year	cost	output	fuel_price	load
1	1970	1140640	0.952757	106650	0.534487
1	1971	1215690	0.986757	110307	0.532328
1	1972	1309570	1.091980	110574	0.547736
1	1973	1511530	1.175780	121974	0.540846
1	1974	1676730	1.160170	196606	0.591167
1	1975	1823740	1.173760	265609	0.575417

Analysis

We will explore this dataset with various plots in attempt to reveal the hidden correlations between the variables. None statistical modelling or null hypothesis testing will be conducted in this report since it is outside the scope of this course of data visualization and of this particular project. However we will try to reveal patterns in the plots and comment on the findings.

Line Graph with Bubble

Firstly, we will try to study the development of the different airlines across these 15 years, in terms of total cost vs fuel price and output index vs load capacity. For this, we will construct a line graph of The total costs vs total output and group the data by each airline. To add another dimension to the graph, we will add bubble to it.

In both plot A and B, it is clear that through 1970-1984 the total costs and output index for all the 6 airlines are increasing in general. Airlines with ID 1 and 2 clearly have much higher costs, output and also the increase of the both through years are much greater in comparison to the rest of 4 airlines.

In plot A, it is easy to see that the fuel price is increasing steadily through years and it has a strong positive correlation to the total costs for all these 6 airlines. In plot B the load capacity for airline 3 seems very stable and rather large through years, but as for the rest of airlines, it varies over time, but it reached its peak around 1979 for most of the airlines which is also the peak for output index. This might indicate a positive correlation between load capacity and output index.

In plot B, the output index increased in the early stage during 1970-1979 and reached its peak around 1979, then it took a little down turn around 1979-1981 and reached to its bottom around 1981, then picked up gradually again 1981 - 1984. For airline with ID 1 this pattern is mostly visible since it is the biggest airline with the highest load capacity and total cost among them all.

Looking back at plot A, it is clear that the fuel price reached its peak around 1981-1982 and this is probably correlated to the lowest output index around the same time in plot B. From this finding, we could say that the higher fuel price around 1978-1980 might have alerted the airlines to cut down their business in scale since the load

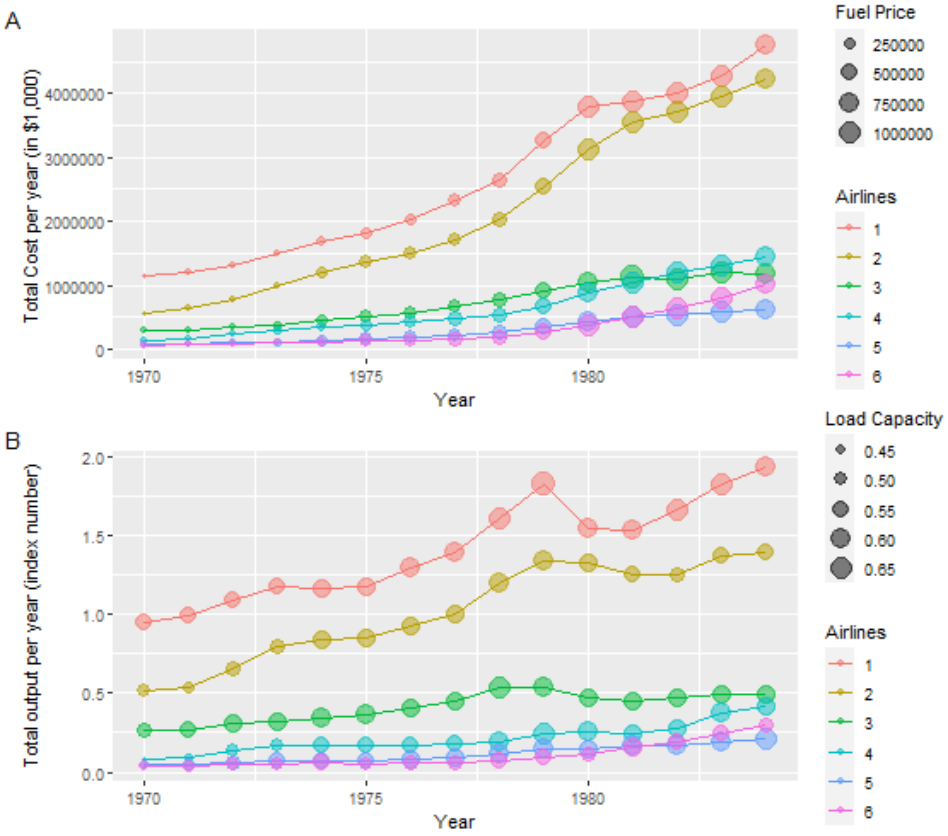

Figure 1: A: Line Graph of total costs for 6 airlines in US between 1970 and 1984 (grouped by airlines, with fuel price as bubble). B: Line Graph of total ouput (index number) for 6 airlines in US between 1970 and 1984 (grouped by airlines, with load capacity as bubble)

capacity decreased and the output index have also declined for the larger airlines 1, 2 & 3 during 1979-1981. Further that the increase of the total costs and fuel price very much negatively affected the profit for the airlines especially for large airlines like 1 & 2 in 1979-1981.

Box Plot

Since we have already explored the numerical variables in the dataset in terms of total cost and profit (output index), now we will dig deeper in these indicators to reveal more patterns. To make the plot more insightful, we will add a new variable *output efficiency* based on

$$outputefficiency = \frac{output}{totalcost} * 1000000$$

across all these 15 years for all the airlines and plot it in a box plot to compare its median and IQR among them.

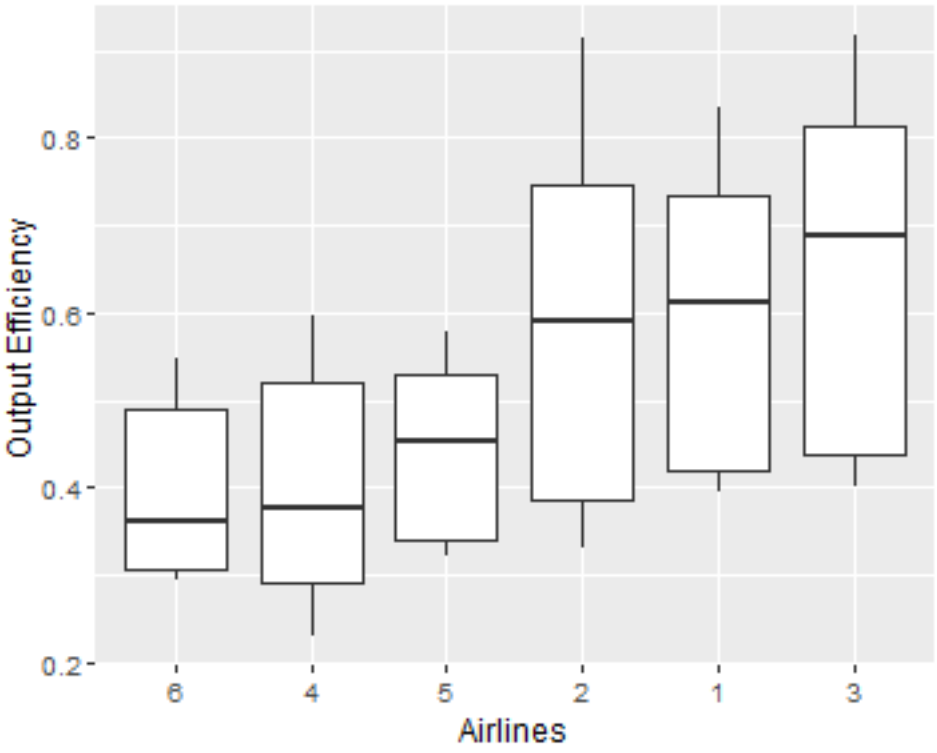

Figure 2: Ordered Box Plot of the Output Efficiency (per 1 million passangers per mile / Total Cost in thousand dollar) for the sex airlines between 1970-1984 in USA.

Looking at this box plot, first we compare the median, it is obvious that airline 3 has the highest output efficiency among all the 6 airlines although it has much lower absolute output index than airline 1 and 2. This might be due to their operating strategy: to operate at low costs (both 1 and 2 have much higher costs, see A in plot 1) and remain high load capacity steadily over years (see B in plot 1). It is clear that

airlines 1, 2 and 3 have much higher output efficiency than the other smaller airlines 4, 5 and 6 in general comparing the median.

Now looking at the IQR, another insight is revealed that operating as larger airlines as 1, 2 and 3 with much larger load capacity, the output efficiency fluctuate much more than those smaller airlines 4, 5 and 6. This indicates that operating large airlines with high load capacity tend to yield higher output efficiency (cost efficient), but also expose output efficiency to more volatility that might be caused by many factors such as fuel price increase or economical recession etc.

Interactive Scatter Plot with Plotly

Now let's explore the correlation between the *fuel price* and *output efficiency* with a scatter plot. The plot will be provided in a separate file uploaded to my github for viewing.

As to our expectation, the output efficiency has a strong negative correlation to the fuel price for all the 6 airlines across 1970-1984 and the correlation is almost linear. This scatter plot confirms again that in spite of the different fuel prices, airline 3 always managed to maintain the highest output efficiency among all airlines. Another insight is that output efficiency for larger airlines 1, 2 and 3 are more negatively affected by the higher fuel price than other 3 smaller airlines, due to their steeper slopes in the plot.

Animation

In the end, here is an animated version of bubble plot (A in plot 1) for the variable *total cost* with *fuel price* coded as the size of the bubble for these six airlines for year 1970-1984, mostly for fun. The plot will be provided in a separate file in format of gif.

Conclusion

This is a rather small dataset (90 observations) with limited variables (6 variables). However with our various plots we managed to reveal the following.

1. Fuel price has a positive correlation to the total cost yet negative correlation to the output.
2. The output increased rather visibly for the larger airlines 1 and 2, but it came with a price since both of them also had much higher total costs than others.
3. Larger airlines in general have higher output efficiency than smaller airlines, but also more sensitive to many factors.
4. Remaining stable high load capacity yet operating with low costs seems to be a winning strategy for high output efficiency for airlines in general.
5. The fuel price keeps increasing over time and so is the same with other operating costs for the airlines. It is crucial for the airlines to adapt its strategy to optimize its business and operation in all aspects to pursuit a high output efficiency.