Business information systems assignment “Case 5”

Damiano Rousselet – 12225451

Report to COO

Introduction

FlyUIBK has ranked below its major client LFUDelta Airlines (LDA) in the EU Aviation Agency September flight delay rankings. LDA has warned FlyUIBK that they must maintain comparable on-time performance or risk losing them as a client. Analysis of data from September, consisting of flight information over 3 airports, will be used to explain the delay and suggest improvements. The report will focus on metrics such as percentage of delayed flights, percentage of on-time flights, and average arrival delay in minutes. We will try to identify points of improvement and give a final suggestion.

[1]

We will analyze the data using the R software. The commands are attached to be replicated and allow further analysis, if necessary. If executed in the given order, with the attachments in the root folder, they will work.

Import and data wrangling

The first step is to import the data correctly and clean it from errors that would hinder our analysis. We use the `read.csv` function with the separator “;”, the headers option to read the first row as headers and the `stringsAsFactors` parameter to set most variables as factors. The data is not perfectly clean, some flights were cancelled, and their delay indicator was replaced by “N//A”, standing for “not available”. Having a non-numeric value is a problem when trying to calculate measurements such as the mean. The NA’s are found for 3 flights operated by LDA, that were cancelled, in the columns “`Arrival.delay.in.minutes`” and “`Delay.indicator`”. This already informs us that LDA cancelled 3 flights, while FlyUIBK did not. Therefore, on the measurement of cancelled flights FlyUIBK performs better. Since they are only 3 entries out of 120 LDA flights the effect of excluding them will be negligable.

Additionally, the "`Number.of.passengers`" column for LDA, contains only not available values, as the company does not provide that data on their flights. The original "N//A"s needs to be converted into the standard used by the software ("NA"), by specifying how they are defined in the dataset. Not having these values makes the data less informative, especially for the number of passengers, which is completely missing for LDA, hence comparisons on that measure will be difficult. One way to possibly expand this analysis would be to infer LDA's number of passengers by different methods of interpolation or imputation. Although, as will be proven in this report this correlation is neither overwhelmingly strong nor practically useful.

We first get an overview of the data:

```
'data.frame':   370 obs. of  14 variables:
 $ Airline                 : Factor w/ 3 levels "","FlyUIBK","LDA": 2 2 2 2 2 2 2 2 2 2 ...
 $ Origin.airport          : Factor w/ 4 levels "","BER","OSL",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Destination.airport     : Factor w/ 4 levels "","BER","OSL",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ Departure.date          : Factor w/ 31 levels "","01.09.2022",..: 2 2 2 3 3 3 4 4 4 5 ...
 $ Scheduled.departure.time: Factor w/ 13 levels "","13:10","13:35",..: 13 2 6 13 2 6 13 2 6 13 ...
 $ Scheduled.arrival.time  : Factor w/ 13 levels "","10:40","14:40",..: 2 3 7 2 3 7 2 3 7 2 ...
 $ Actual.arrival.time     : Factor w/ 200 levels "","10:38","10:40",..: 9 32 98 12 35 102 3 24 90 11 ...
 $ Arrival.delay.in.minutes: int  10 10 9 13 13 13 0 0 0 12 ...
 $ Delay.indicator         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Day.of.Week             : int  5 5 5 6 6 6 7 7 7 1 ...
 $ Route.Code              : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Number.of.passengers    : int  100 94 92 182 193 192 104 95 101 84 ...
 $ X                       : logi  NA NA NA NA NA NA ...
 $ X.1                     : logi  NA NA NA NA NA NA ...
```

Table 1 – Dataset classes and columns

We notice that two columns have been added and automatically named "X" and "X.1". This is a mistake; they must be removed. Additionally, when we load the entire dataset, we notice from the 370 observations that 10 empty lines at the end of the data were created, they also must be removed. Further, we set the `Departure.date` variable to Class Date, to better work with it.

For readability, we give the conventional weekday names to the values representing them in the `Day.of.Week` column. We do the same for routes and the delay indicator. Before changing `Delay.indicator`, we make a copy with the indicator as an integer, which we will use for calculations.

Data analysis

When filtering the data by `Route.Code` and doing the difference between `Scheduled.arrival.time` and `Scheduled.departure.time`, we find that the two airlines schedule different flight durations for the same routes. This significantly influences the minutes of delay since they are computed as the difference between the actual arrival time and the scheduled arrival time. These are the different scheduled flight durations:

	BER/VIE	VIE/BER	VIE/OSL	OSL/VIE
FlyUIBK	1:30	1:30	1:10	1:10
LDA	1:40	1:40	1:15	1:15

Table 2 – Scheduled flight durations

LDA schedules 10 minutes longer for route BER/VIE - VIE/BER, and 5 minutes longer for VIE/OSL – OSL/VIE. This means that the delay of FlyUIBK has to be adjusted by reducing it by the longer flight duration of LDA. After preparing the data in Excel, we import it into R and add it as columns to the `data.frame`, with names `Adjusted.delay` and `Adjusted.indicator`.

We get a synthesis of the data with the most relevant columns:

```
    Airline     Origin.airport Destination.airport Arrival.delay.in.minutes Delay.indicator    Day.of.Week    Route.Code   Number.of.passengers
       :  0        :  0            :  0            Min.    :-13.00          On Time:263      Friday   :60   BER/VIE:120   Min.   : 34.0
 FlyUIBK:240    BER:120         BER:120            1st Qu.:  5.00          Delayed: 94      Monday   :48   VIE/BER:120   1st Qu.: 85.0
 LDA    :120    OSL: 60         OSL: 60            Median : 11.00          NA's   :  3      Saturday :48   VIE/OSL: 60   Median :104.0
                VIE:180         VIE:180            Mean   : 14.11                           Sunday   :48   OSL/VIE: 60   Mean   :112.9
                                                   3rd Qu.: 15.00                           Thursday :60                 3rd Qu.:128.0
                                                   Max.   :153.00                           Tuesday  :48                 Max.   :204.0
                                                   NA's   :3                                Wednesday:48                 NA's   :120
 Adjusted.delay     Adjusted.indicator
 Min.   :-17.000    On Time:307
 1st Qu.: -2.000    Delayed: 50
 Median :  4.000    NA's   :  3
 Mean   :  8.227
 3rd Qu.: 13.000
 Max.   :143.000
 NA's   :3
```

Table 3 – Descriptive statistics

From `Airline`, we notice that FlyUIBK has 240 flights and LDA 120. From `Origin.airport` and `Destination.airport`, we see that most flights (180) are to/from Vienna, followed by Berlin (120) and Oslo (60), the most flown route is BER/VIE and vice versa (120), followed by VIE/OSL with 60 flights in each way.

From `Arrival.delay.in.minutes`, we notice that the mean is 14 minutes, but from `Adjusted.delay` we notice that the mean is 8.2 minutes.

Further, we have to consider that some flights arrived earlier and the sum of all these anticipated arrivals will skew the mean, by reducing the actual minutes of delay.
We calculate the mean excluding early arrivals and get:

$$Original\ mean\ delay = 15.40841$$
$$Adjusted\ mean\ delay = 9.432432$$

Note: `Adjusted delay excluding anticipated flights` removes flights that arrived earlier before the time adjustment, then adds the extra time, otherwise the time adjustment would lead to more flights being marked as "early" and hence removed, leading to an even longer average delay.

Summarizing the different mean delays:

	Mean delay min. With early arrivals	Mean delay min. w/o early arrivals
Original delay	14	15.4
Adjusted delay	8.2	9.4

Table 4 – Average delays for adjusted flight durations

The adjusted mean delay is much lower than the original, and obviously, both mean delays increase if we exclude early arrivals. Hence, the different scheduled flight time has a strong influence on delay minutes, favoring LDA. The mean delay of our company drops sharply if we adjust to flight duration.
Further, From `Delay.indicator` we see that 3 flights were cancelled and can verify with:

```
> filter(complete_dataset, Actual.arrival.time == "Cancelled")
  Airline Origin.airport Destination.airport Departure.date Scheduled.departure.time
1     LDA            BER                 VIE     2020-09-09                    21:55
2     LDA            BER                 VIE     2020-09-26                    21:55
3     LDA            VIE                 BER     2020-09-09                    13:35
```

Table 5 – Cancelled flights

LDA had 3 of their 120 flights cancelled, which is 2.5% of all their flights. Two are on the same route and leave at the same time (21:55), but on different days.

By dividing $\frac{Delayed}{Total\ flights - Cancelled} = \frac{94}{360-3} = 0.2633 \rightarrow 26\%$ the percentage of delayed flights is obtained.

The `Adjusted.indicator`, gives us $\frac{50}{360-3} = 0.1400 \rightarrow 14\%$, which is almost half of the unadjusted delay.

The busiest days are Friday and Thursday with 60 flights each, followed by all other days with 48 flights.

Finally, from `Number.of.passengers`, we can see that the mean is 113, this is only available for FlyUIBK.

To get insight about the performance, we calculate: average of normal and adjusted delay, their standard deviation, including or excluding early arrivals (not for the delay indicator as that would not make sense):

```
overall_performance
avg_delay    StDev avg_adj_delay StDev_adj pct_delayed pct_delayed_adjusted
 14.10924 23.05774      8.226891   22.9139   0.2633053             0.140056
overall_performance_early
avg_delay    StDev avg_adj_delay StDev_adj pct_delayed pct_delayed_adjusted
 15.40841 23.33016      9.432432  25.32333   0.2633053             0.140056
```

Table 6 Performance indicators overall

Note: `overall_performance_early` does not contain early arrivals.

Now, we will look at the percentage of flights that are delayed for each airline and compare them. This will highlight which airline is experiencing more issues.

We need to use the dataset with `Delay.indicator` as an integer to compute the mean.

We compare the performance on the average unadjusted and adjusted delay, including or excluding early arrivals (not for delay indicator = “pct_delayed”):

```
Fly_UIBK_performance
avg_delay avg_adj_delay pct_delayed pct_delayed_adjusted
  15.6625        6.9125      0.2625           0.07916667
Fly_UIBK_performance_early
avg_delay avg_adj_delay pct_delayed pct_delayed_adjusted
 16.86283      8.057522      0.2625           0.07916667
LDA_performance
avg_delay avg_adj_delay pct_delayed pct_delayed_adjusted
 10.92308      10.92308   0.2649573            0.2649573
LDA_performance_early
avg_delay avg_adj_delay pct_delayed pct_delayed_adjusted
 12.33645      12.33645   0.2649573            0.2649573
```

Table 7 – Performance indicators per airline

The percentage of delays is close: 26.2% for FlyUIBK and 26.5 for LDA. The larger difference lies in the average delay minutes, which are higher for FlyUIBK, no matter if we consider or exclude early arrivals. Indeed, if we exclude early arrivals the average delay is 36.7 % higher for FlyUIBK and 43% higher if we include them, so LDA has more early arrivals than FlyUIBK.

On the other hand, if we consider the average adjusted delay, FlyUIBK has a better performance than LDA, whether we include early arrivals or not. The adjusted delay indicator drops from 26% to 8% for FlyUIBK, 3.25 times less than originally thought. The average adjusted delay minutes are now larger for LDA: if we include early arrivals 59% larger than FlyUIBK, and 53% larger if we exclude early arrivals.

We calculate the average arrival delay, the adjusted delay, and the percentage of flights that were delayed for each day of the week, excluding early arrivals:

Note: LDA does not have an adjusted delay as only FlyUIBK delays were adjusted.

Summarizing, we get:

Day.of.Week	avg_delay	avg_adj_delay	avg_delay_FlyUIBK	avg_adj_delay_FlyUIBK	avg_delay_LDA
Friday	29.8	23.7	35.7	27	16.6
Monday	33.3	27.4	41.4	32.7	16.1
Saturday	6.58	0.651	5.17	-3.62	9.5
Sunday	12.2	6.42	12.1	3.31	12.6
Thursday	7.32	1.26	6.05	-2.79	10.1
Tuesday	6.69	1	5	-3.91	9.69
Wednesday	7.02	0.773	5.48	-3.39	10.7

Table 8 – Average delays per day

Percentage delayed:

Day.of.Week	%delayed	%delayed_adjust	%delayed_FlyUIBK	%delayed_adjust_FlyU	%delayed_LDA
Friday	0.828	0.414	0.8	0.2	0.889
Monday	0.894	0.511	0.906	0.344	0.867
Saturday	0	0	0	0	0
Sunday	0.0833	0.0417	0.0625	0	0.125
Thursday	0	0	0	0	0
Tuesday	0	0	0	0	0
Wednesday	0	0	0	0	0

Table 9 - Percentages of delays per day

Overall, Monday is the day with the highest average delay: over 33 minutes. It also has the highest percentage of delays with 89% of flights being delayed at least 15 minutes. Other long delays occur around the weekend, except for Saturday.

FlyUIBK has very large delays on Mondays (90% of all daily flights) with an average delay of 41.4 minutes, and on Fridays. On other days it is somewhat better than LDA. The percentage of delayed flights is similar amongst both airlines. The worst average delay is much higher for FlyUibk (41.4$_{min}$), almost 2.5 times that of LDA (16.6$_{min}$).

If we look at the adjusted delay, there is a sharp drop especially on Tuesday, Wednesday, Thursday, and Saturday. On Sundays, FlyUIBK also experiences sharp reductions in delays, although not early arrivals on average. FlyUIBK has shorter delays than LDA on Tuesday, Wednesday, Thursday, Saturday and Sunday. Only on Mondays and Fridays, LDA has shorter delays.

LDA has most flights on Friday (89%) delayed, with an average delay of 16.6 minutes.

This confirms the aggregate data from above, where FlyUIBK had overall fewer delays than LDA, all days combined.

To analyze the relationship between "`Day.of.Week`" and "`Arrival.delay.in.minutes`", we use a scatterplot, for the days with the highest delays for FlyUIBK: Friday and Monday.

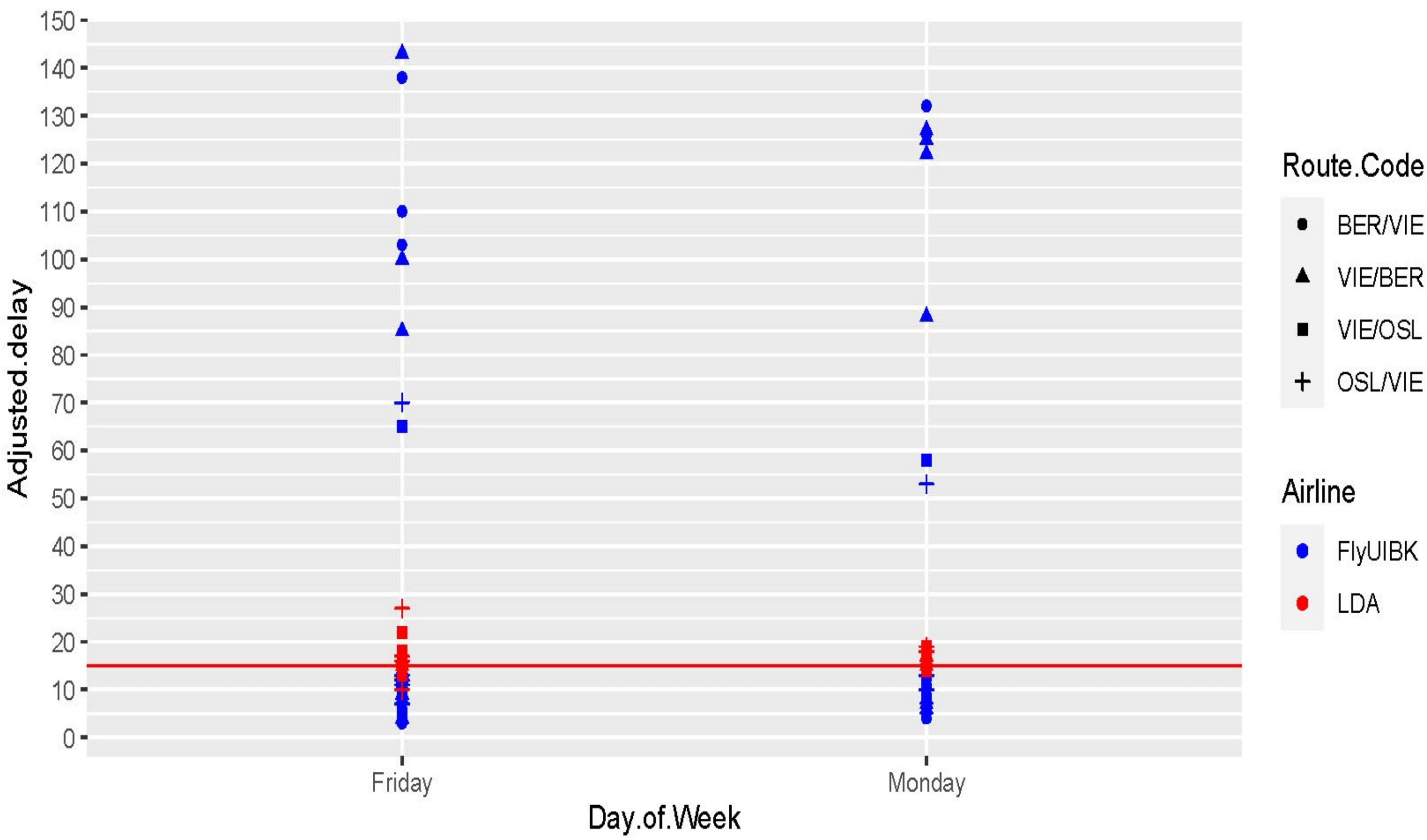

Figure 1 – Adjusted delays per route, airline, Friday, Monday

There are larger outliers on Mondays and Fridays, from FlyUIBK, as seen by the colour. The red line represents the 15-minute-delay threshold. From the shape, we can see that the longest delays are on both ways of the VIE/BER route, which is coherent with the information about airport congestion.

It is also the case that FlyUIBK operates 3 daily flights between VIE and BER and LDA one, for a total of 4. On the OSL/VIE route, there are only 2 daily flights, one from each carrier. Therefore, it is to be expected that the total number of delays on the first route is larger than the second, in absolute terms, because there are twice as many flights. To get the relative delays they must be divided by the number of flights per route.

There are 2 outliers for FlyUIBK flights leaving OSL to VIE, which is surprising as Oslo is a low-congestion airport. One reason could be that those flights had to queue for landing in Vienna as it is a medium-congestion airport.

There are no early arrivals on Mondays or Fridays.

We verify in more detail the average delays per route.

Route	Average delay	Average adjusted delay
BER/VIE	16.3	8.7
VIE/BER	18	10.5
VIE/OSL	11.8	9.16
OSL/VIE	11.5	8.88

Table 10 - Average delays per route

We see that the BER/VIE – VIE/BER route is still the one with longer delays even relative to the number of flights on that route. Although, with the adjusted flight duration, the differences become much smaller, another indication that the scheduled flight duration has a large impact on delays.

The delays are not extremely different per route, suggesting that the route does not have a strong effect on the minutes of delay.

Correlation testing

To test whether the number of passengers has an influence on the delays, we first need to know if the data is normally distributed. If it is not, a non-parametric test is needed to test for correlation.

We will perform the Shapiro-Wilk test of normality on the `Arrival.delay.in.minutes` column in the `complete_dataset`. The test returns a p-value, which can be interpreted as follows:

- If it is less than 0.05, reject the null hypothesis that the data is normally distributed, and conclude that it is significantly non-normal.

- If it is greater than or equal to 0.05, we cannot reject the null hypothesis: the data may be considered normal.

```
        Shapiro-Wilk normality test

data:  complete_dataset$Arrival.delay.in.minutes
W = 0.47132, p-value < 2.2e-16
```

Finding 1 - Normality of delay minutes distribution

The p-value of the Shapiro-Wilk test is less than 2.2e-16, which suggests that the data is not normally distributed. Therefore, it is not appropriate to use a parametric test. Instead, we use a nonparametric test such as Spearman's rank correlation test, which does not assume normality. We do this with the `cor.test` function with the argument set to "`spearman`". We add the `exact = F` parameter because there are some data-points with the same values, which thus cannot be ranked by Sparkman's test.

```
        Spearman's rank correlation rho

data:  complete_dataset$Adjusted.delay and complete_dataset$Number.of.passengers
S = 1136601, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5066749
```

Finding 2 - Correlation between Number of passengers - Adjusted delay

The p-value of less than 2.2e-16, being much smaller than 0.05, indicates that the correlation between the `Number.of.passengers` and the `Adjusted.delay` is statistically significant, even if `exact = F`, as the difference to 0.05 is very large. This implies that it is unlikely to have occurred by chance. The sample estimate of the correlation coefficient is 0.50667, which indicates a moderate positive correlation between the two variables. Therefore, as the number of passengers increases, the adjusted delay tends to increase as well. Overall, this suggests that the number of passengers may be related to the arrival delay, but other factors may also be contributing to the delays, for example, the day of the week (Monday, Friday) or the route taken (BER/VIE and vice versa).
Further, it is important to keep in mind that correlation does not necessarily imply causation, it could be that the routes with longer delays happen to have more passengers than other routes, but that the delay is explained more by the route (or another factor), than by the number of passengers.

Since adjusting the minutes of delay is done by subtracting 2 constants from the observations, it is a linear transformation, which will result in translating the distribution along the x-axis and a change to the standard deviation. The shape of the distribution will still be non-normal. To test if there is a statistical relation between `Day.of.Week` and `Adjusted.delay`, we use another test, because Spearman's test is not for categorical variables, which `Day.of.Week` is.

We use a non-parametric test: the Kruskal-Wallis test. This test can be performed using `kruskal.test`. It ranks the data from all the days and then compares the ranks of the observations within each group. The test statistic is based on the sample sizes of the groups and the ranks of the observations.

The H_0 is that the medians of all groups are equal. The H_1 is that at least one of the medians is different.

```
        Kruskal-Wallis rank sum test

data:  Adjusted.delay by Day.of.Week
Kruskal-Wallis chi-squared = 160.08, df = 6, p-value < 2.2e-16
```

Finding 3 - Correlation between day of week - Adjusted delay

As the p-value is much smaller than 0.05, and even than 0.01, there is strong evidence against the null hypothesis, of all medians being equal.

For the relation between `Route.Code` and `Adjusted.delay`, we use the same test:

```
        Kruskal-Wallis rank sum test

data:  Adjusted.delay by Route.Code
Kruskal-Wallis chi-squared = 4.2263, df = 3, p-value = 0.238
```

Finding 4 - Correlation between route - Adjusted delay

The p-value is much larger than 0.05. The correlation is not statistically significant, it is not likely to be a meaningful relationship. This confirms what was already suggested by the table of `Adjusted.delay` per route.

An important detail about the data is that we only have a sample from September, hence predictions about the entire "population" of yearly flights will be skewed. Flight data varies substantially across months because it is strongly impacted by seasonal events like holidays, weather, and politics. The sample is not representative of the population.

It would be possible to use a model-based approach, such as multiple linear regression, to predict the values for the rest of the year based on September, and then compare the predicted values to the actual values (from September) to assess the accuracy of the model. However, the accuracy of these predictions would depend on the quality and relevance of the sample data and the assumptions made in the model.

For the reasons explained above, the predictions would be quite useless for generalizing to an entire year. If the sample would for example be 300 random flights from the year, a model would make sense.

Interpretation

The most important finding is that LDA schedules longer flight times than FlyUIBK. This finding can be generalized to the entire year, assuming it is standard practice and not only a coincidence during this month.
It is possible to see that if the flight duration of FlyUIBK is adjusted to match LDA's, the company has a shorter mean delay, while LDA has up to 59% longer delays.
However, on Fridays and Mondays FlyUIBK has the worst delays, much more than LDA.
The route has some influence on delays, but after adjusting the flight durations, the differences become severely smaller.
Finally, the number of passengers is related to a delay in arrival time, even when adjusting the flight durations. This is to be expected as more passengers take more time to board and leave.

The best suggestion is for FlyUIBK to adjust its flight durations to LDA. This would be a financially convenient solution, with an instant effect – a low-hanging fruit.
Since FlyUIBK is being judged by LDA for its poor performance, these findings prove that the real performance is actually better than LDA's.
An additional idea is to reduce flights on Fridays and Mondays when most delays happen, or to increase staff and optimize procedures. This will be more complicated to implement, and a cost-benefit analysis should be done beforehand.
Changing routes would be highly unpractical and yield only small improvement, at a high cost, and should thus be avoided. These delays reflect the fact that some airports have higher congestion, which is not under FlyUIBK's influence.
The number of passengers is related to the delays, a possible suggestion would be to increase staff and boarding procedures on crowded flights.

Sources:

- [1] Business Information Systems: University of Innsbruck - Case: On-Time Performance of FlyUIBK
- Siegel, Andrew F - Practical business statistics-Academic Press, Elsevier (2016)
- Ross S.M. - Introductory statistics Elsevier academic press (2017)
- Data analysis and graphics using R: an example-based approach / John Maindonald, W. John Braun. – 3rd ed
- https://www.r-project.org/help.html
- 2011 Mar-Apr; 59(2): 85–86. doi: 10.4103/0301-4738.77005 PMCID: PMC3116565 PMID: 21350275 "How to choose the right statistical test?" - Barun K Nayak and Avijit Hazra1