BIỂU DIỄN THÔNG TIN

đồ ÁN MÔN HỌC: CARGO

nHÓM 9

2017

**Nội dung**

[I. Thông tin nhóm 2](#_Toc503788639)

[II. Kế hoạch thực hiện 3](#_Toc503788640)

[1. Kế hoạch tổng quan 3](#_Toc503788641)

[2. Kế hoạch chi tiết cho từng giai đoạn: 3](#_Toc503788642)

[III. Tổng quan về dữ liệu 4](#_Toc503788643)

[IV. Công trình nghiên cứu liên quan 6](#_Toc503788644)

[1. Link tài liệu tham khảo: 6](#_Toc503788645)

[2. Tóm tắt nội dung: 6](#_Toc503788646)

[V. Trực quan hóa dữ liệu 10](#_Toc503788647)

[1. Mục tiêu: 10](#_Toc503788648)

[2. Phân tích dữ liệu 10](#_Toc503788649)

[Câu hỏi 1: 10](#_Toc503788650)

[Câu hỏi 2: 12](#_Toc503788651)

[VI. Cài đặt 15](#_Toc503788652)

[VII. Nhận xét và tự đánh giá 19](#_Toc503788653)

# Thông tin nhóm

|  |  |  |  |
| --- | --- | --- | --- |
| STT | MSSV | Họ tên thành viên | Vai trò |
| 1 | 1412503 | Nguyễn Thị Thanh Thảo | Thành viên |
| 2 | 1412542 | Nguyễn Hà Tiên | Nhóm trưởng |
| 3 | 1412543 | Nguyễn Thủy Tiên | Thư ký |
| 4 | 1412595 | Võ Thị Thanh Trúc | Thành viên |

# Kế hoạch thực hiện

## Kế hoạch tổng quan

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Giai đoạn | Công việc | Thời gian |
| 1 | Phân tích dữ liệu | Phân tích thông tin, nội dung bộ dữ liệu | 06/10/2017 → 02/11/2017 |
| Tìm hiểu các công trình liên quan |
| Tìm hiểu các công cụ hỗ trợ trực quan hóa |
| Xác định mục tiêu của trực quan hóa, đặt câu hỏi để phân tích, thực hiện mục tiêu |
| Tìm hiểu các công cụ hỗ trợ trực quan hóa |
| Dùng công cụ hỗ trợ để vẽ các trực quan hóa đơn giản, phân tích câu hỏi đã đặt ra |
| Báo cáo giữa kỳ cho giáo viên (03/11) |
| 2 | Thiết kế công cụ trực quan hóa | Phân tích và lựa chọn cách trực quan hóa | 04/11/2017 → 20/11/2017 |
| Dùng công cụ hỗ trợ để phác họa trực quan |
| 3 | Cài đặt công cụ trực quan hóa | Tìm hiểu công cụ hỗ trợ cài đặt (D3, javascript…) | 16/11/2017 → 10/01/2018 |
| Thực hiện cài đặt |
| 4 | Làm báo cáo, poster | Viết báo cáo về quy trình làm việc, sản phẩm của nhóm | 11/012018 → 14/01/2018 |
| Thiết kế poster |

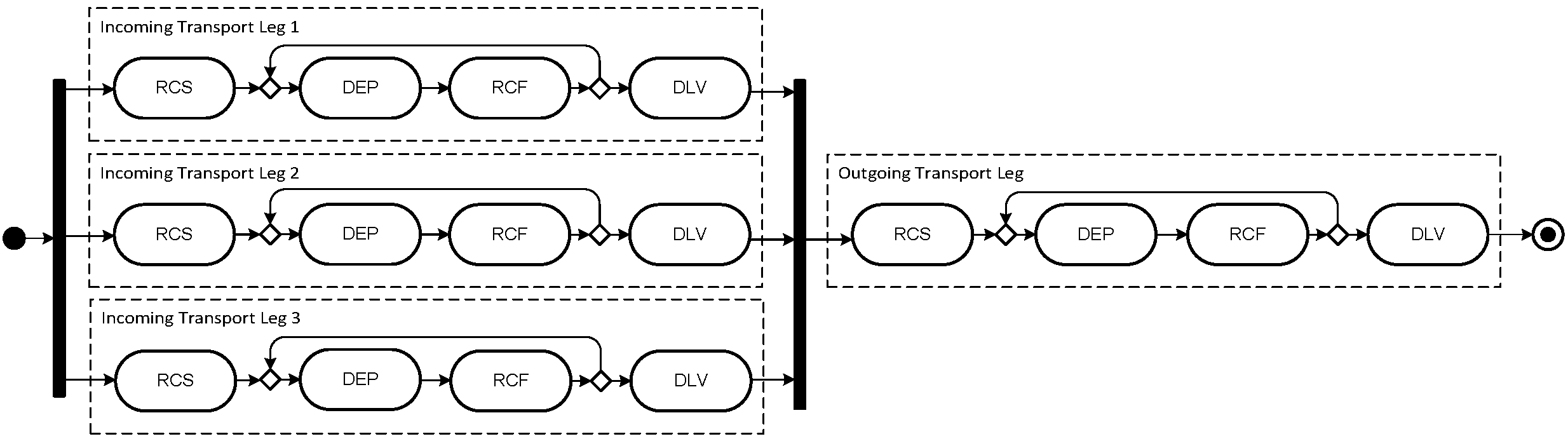
## Kế hoạch chi tiết cho từng giai đoạn:

File Planning.docx

# Tổng quan về dữ liệu

Bộ dữ liệu của nhóm có tên là Cargo, thể hiện sự vận chuyển hàng hóa của các một công ty trong vòng 5 tháng, gồm có 98 thuộc tính và 3943 dòng dữ liệu đại diện cho một quá hoạt động vận chuyển hàng hóa từ nhà cung cấp đến nơi nhận.

Quá trình vận chuyển hàng hóa của công ty được thể hiện như sau:

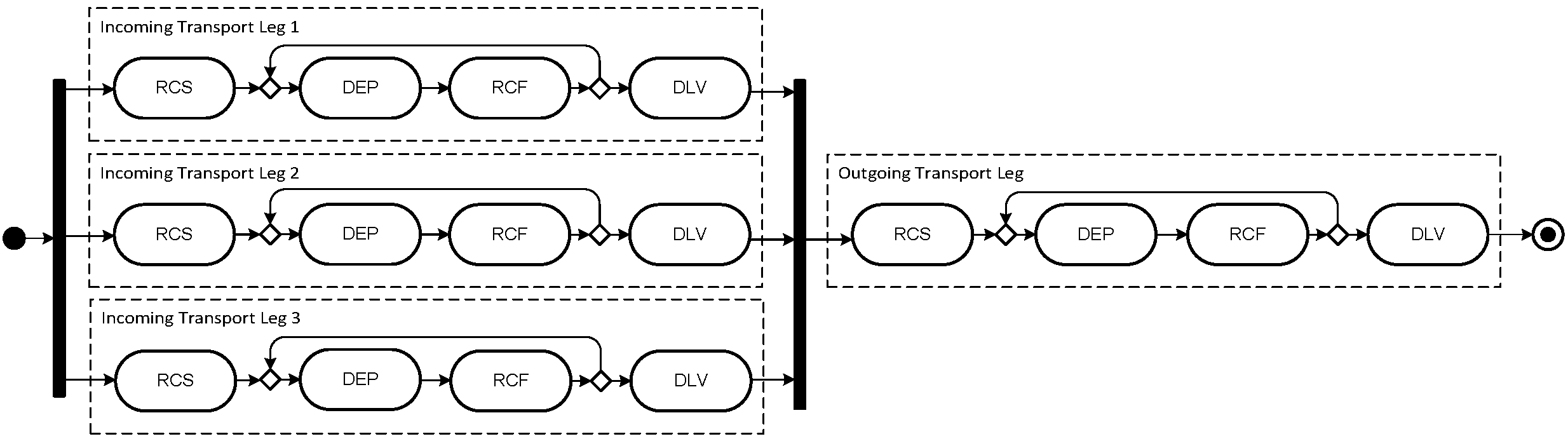


Hình 1: . Sơ đồ hoạt động vận chuyển hàng hóa của công ty Cargo 2000

Một quá trình vận chuyển hàng hóa của công ty được chia các tuyến vận chuyển nhỏ. Trong đó có 3 tuyến vận chuyển đến từ các nhà cung cấp được thực hiện song song và 1 tuyến vận chuyển đi.

Hình 2: Sơ đồ hoạt động của một tuyến vận chuyển

Cụ thể 1 tuyến vận chuyển như sau:

**

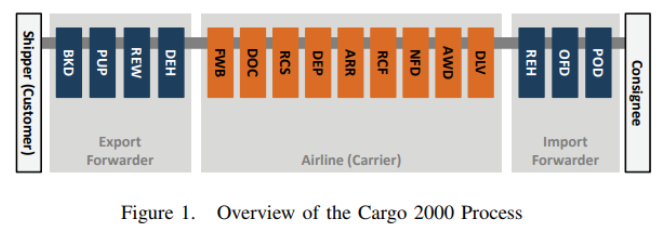
* Một tuyến vận chuyển gồm có 4 dịch vụ RCS, DEP, RCF, DLV.
* RCS: dịch vụ kiểm tra hàng hóa ở sân bay đi
* DEP: dịch vụ vận chuyển hàng hóa lên máy bay
* RCF: dịch vụ vận chuyển hàng hóa đến sân bay đến
* DLV: dịch vụ giao hàng
* Trong một tuyến vận chuyển có thể cần vận chuyển qua nhiều sân bay, máy bay khác nhau. Vì vậy, 2 dịch vụ DEP và RCF có thể được lặp lại cho mỗi lần thực hiện chuyển hàng hóa qua máy bay khác tại sân bay.
* Ở mỗi dịch vụ, dữ liệu gồm có thời gian vận chuyển dự kiến và thời gian thực tế. Dịch vụ DEP và RCF đều được gắn với một địa điểm nhất định.

# Công trình nghiên cứu liên quan

## Link tài liệu tham khảo:

<https://pdfs.semanticscholar.org/86a0/23be6c6ed75a48cfab5afa1cc25b7caf3d45.pdf>

## Tóm tắt nội dung:

Bài nghiên cứu trên sử dụng cùng bộ dữ liệu Cargo 2000 nhưng thể hiện chi tiết hơn.

Quy trình vận chuyển hàng hóa Cargo gồm 3 giai đoạn:

* Giai đoạn 1: Giai đoạn nhận hàng hóa từ khách hàng đưa đến sân bay đi.

BKD: đặt hàng hóa từ shipper

PUP: nhận hàng hóa từ khách hàng.

REW: nhận hàng hóa từ kho hàng.

DEH: xe tải khởi hành đến sân bay.

* Giai đoạn 2: Giai đoạn vận chuyển hàng hóa qua các sân bay. Giai đoạn này ứng với bộ dữ liệu mà chúng em phân tích ở trên.

FWB: giai đoạn lấy hóa đơn.

DOC: xe tải đến sân bay khởi hành.

RCS: kiểm tra hàng hóa tại sân bay khởi hành.

DEP: xác nhận hàng hóa.

ARR: chuyến bay đến điểm đến.

RCF: chấp nhận hàng hóa đến sân bay.

NFD: tất cả hàng hóa và tài liệu sẵn sàng để nhận hàng.

DLV: phân phối hàng hóa đến nơi nhập

* Giai đoạn 3: Giai đoạn vần chuyển hàng hóa từ sân bay đến đến chỗ nhận hàng.

REH: nhận hàng hóa tại kho nhập khẩu.

OFD: giao xong hàng hóa vào kho.

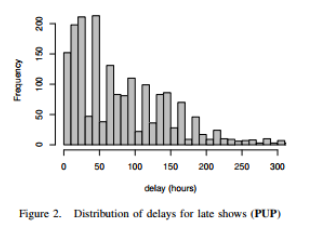
POD: giao hàng cho người nhận hàng.

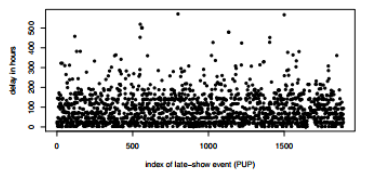
Theo như phân tích, người ta thấy được có 3 nguyên nhân chính dẫn đến giao hàng trễ:

1. Late shows:

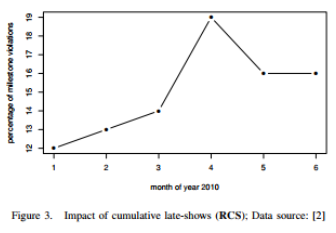
Các trường hợp xảy ra late shows:

* Nhận hàng từ khách hàng để giao trễ (PUP)
  + Quá trình nhận hàng hóa từ kho trễ (REW)
  + Xe tải rời khỏi sân bay trễ (DEH)
  + Trễ do giai đoạn kiểm tra hàng hóa tại sân bay (RCS)

Người ta phân tích được giai đoạn nhận hàng hóa từ khách hàng bị late shows nhiều nhất. Thời gian trễ (h) được thể hiện thông qua 2 biểu đồ bên đưới. Qua 2 biểu đồ trên ta thấy được khoảng thời gian trễ trong khoảng từ 0 -> 500 (giờ), trong đó khoảng thời gian trễ từ 0->50 (giờ) là nhiều nhất.



Người ta còn phần tích một sô nguyên nhất khác dẫn đến giao hàng trễ. Dưới đây là phân tích độ trễ theo từng tháng của năm 2010 của giai đoạn RCS.



Ở biểu đồ trên ta thấy tháng 4 có thời gian trễ nhiều nhất so với các tháng còn lại. Theo như nghiên cứu người ta thấy đươc vào tháng 4/2010 đã xảy ra núi lửa phun trào nên việc vận chuyển hàng hóa bị ảnh hưởng.

1. No shows

Các trường hợp xảy ra no shows:

* Khách hàng không có hàng để giao cho công ty thực hiện việc vận chuyển (BKD)
* Kiểm tra hàng hóa tại sân bay khởi hành thất bại (RCS)

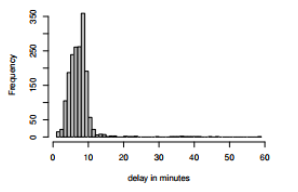
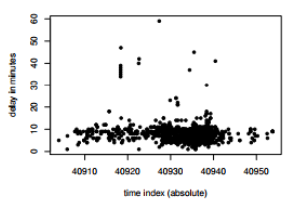
1. Data late

Các trường hợp xảy ra data late:

* Giai đoạn tạo hóa đơn bị trễ (FWB).

Vì dữ liệu được cập nhật liên tục tại hệ thống cục bộ nhưng hệ thống trung tâm lại cập nhật theo định kì -> dữ liệu có thể cập nhật trễ -> việc xử lý dữ liệu của hệ thống bị chậm trễ:

Giai đoạn nhận hóa đơn là giai đoạn xảy ra data late nhiều nhất. Khoảng thời gian trễ của giai đoạn nằm trong khoảng từ 0 -> 60 phút, trong đó 82% trễ trong vòng 10 phút và 18% trễ trong vòng 60 phút



* Thông qua việc phân tích các nguyên nhân gây trễ trên, người ta biết được giai đoạn nào bị trễ do nguyên nhân nào, từ đó có thể dự đoán và tìm cách cải thiện chất lượng của công ty.

# Trực quan hóa dữ liệu

## Mục tiêu:

Nhận biết chất lượng dịch vụ của công ty để có thể cải thiện, nâng cao chất lượng hoạt động của công ty.

## Phân tích dữ liệu

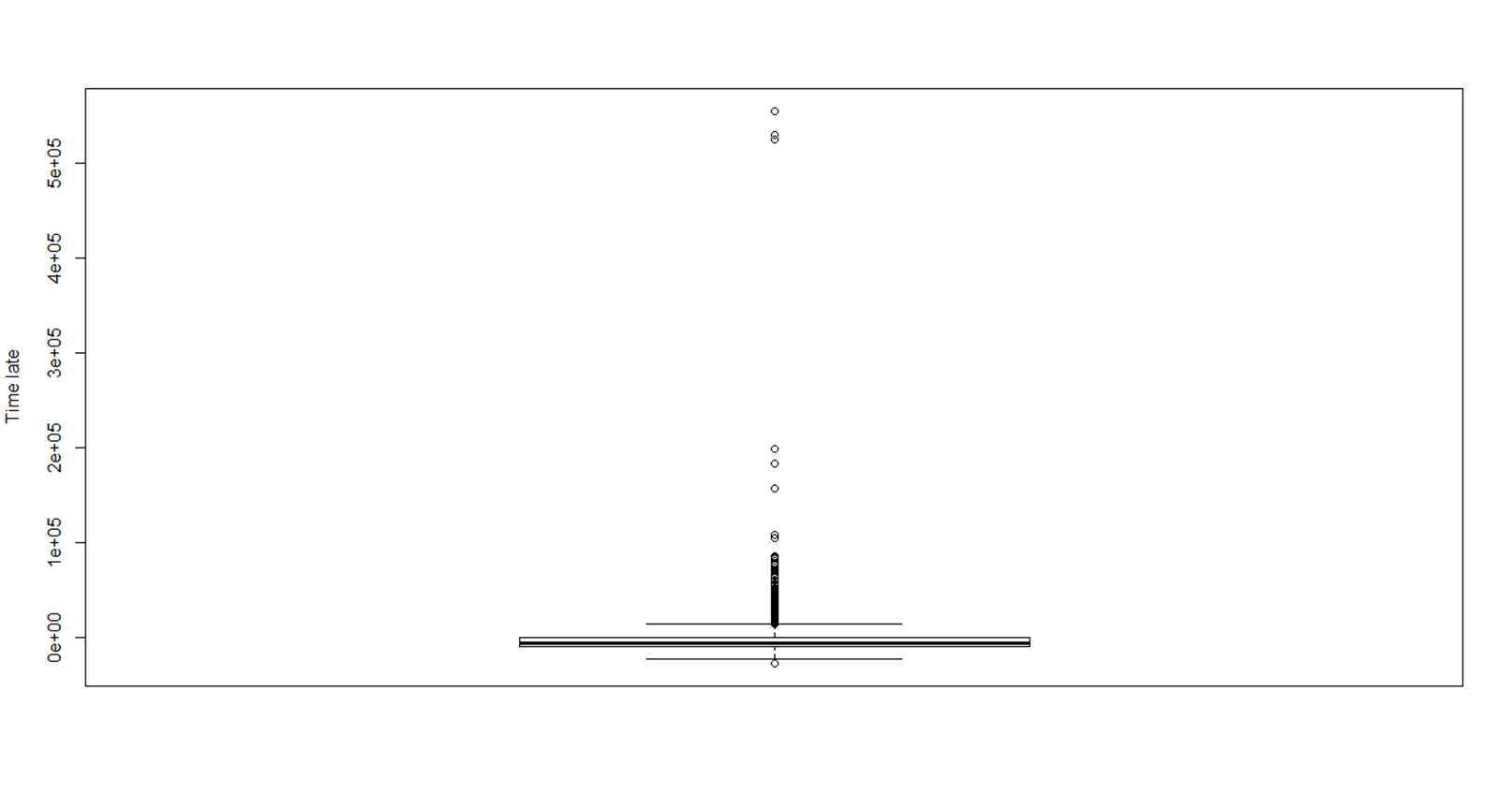
Các câu hỏi nhóm đặt ra cho bộ dữ liệu:

### Câu hỏi 1:

Chất lượng hoạt động của công ty như thế nào?

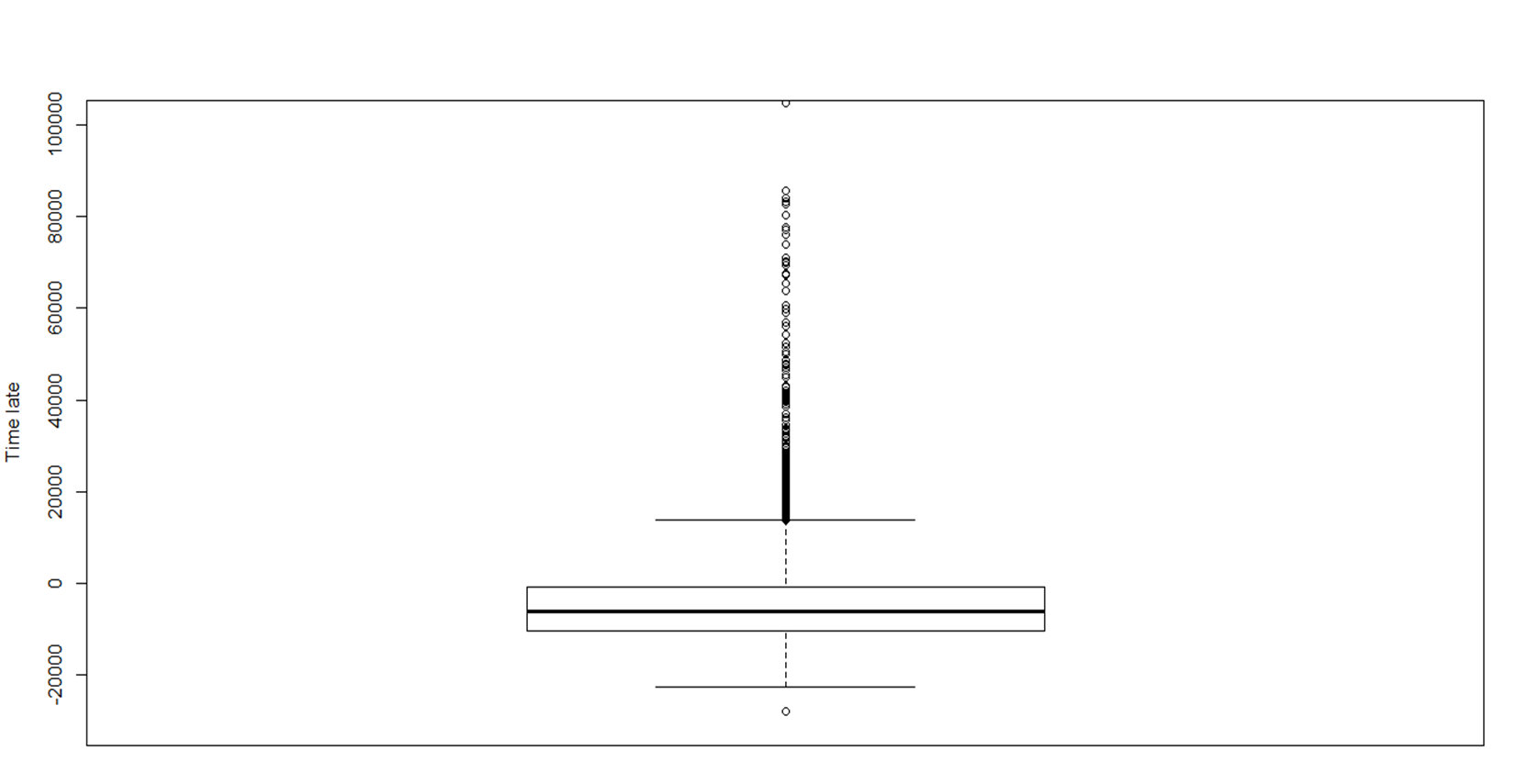
Để trả lời cho câu hỏi này thì ta sẽ tính toán độ trễ (chênh lệch giữa thời gian dự kiến và thời gian thực tế) của từng dịch vụ của từng tuyến sau đó tổng hợp thành toàn bộ ta được độ trễ của toàn bộ quá trình vận chuyển.

Độ trễ của toàn quá trình vận chuyển:



Hình 3: Tổng quan về độ trễ của toàn quá trình vận chuyển hàng hóa của công ty

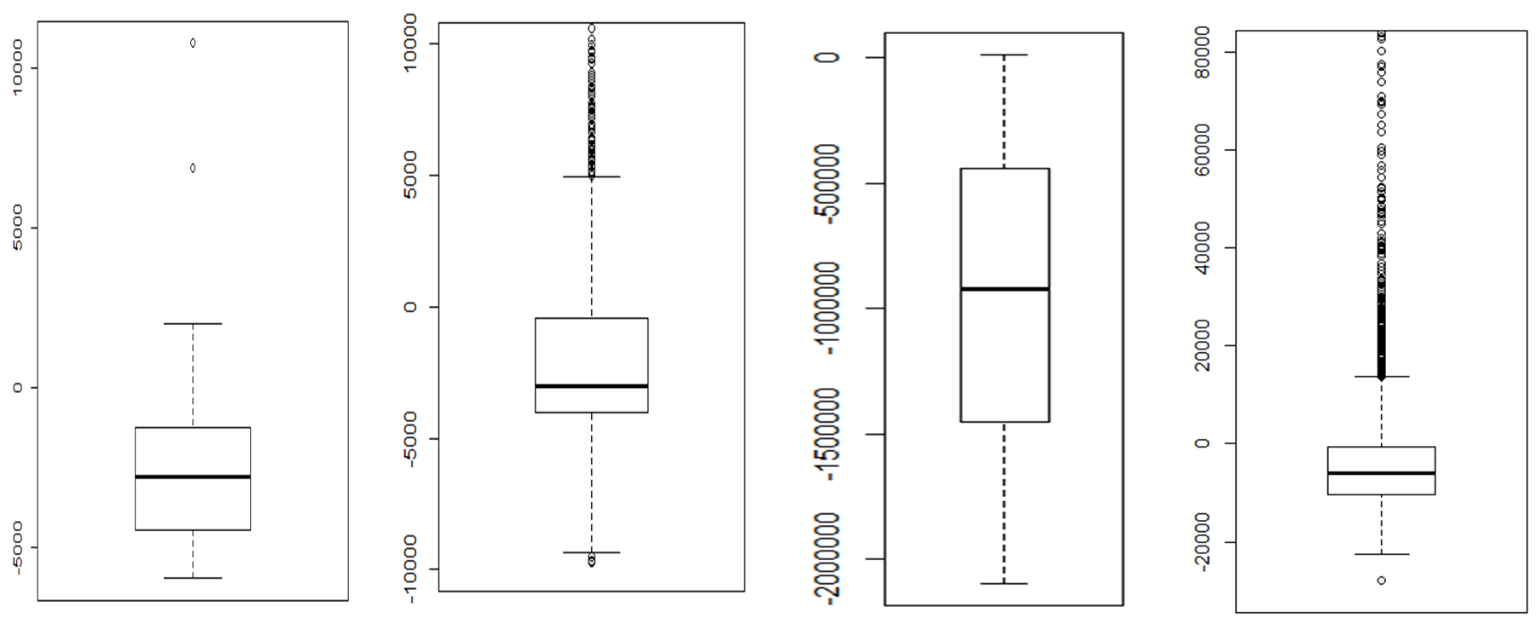
Ở biểu đồ hộp này, các tuyến vận chuyển sớm so với dự định, ngược lại > 0 là các tuyến vận chuyển trễ, từ đó ta thấy khoảng giá trị ngoại lai trễ rất lớn so với khoảng phân vị.



Hình 4: Chi tiết độ trễ của toàn quá trình vận chuyển hàng hóa

Ở biểu đồ này, giảm khoảng giá trị lớn nhất của biểu đồ lại để thấy được chi tiết hơn khoảng phân vị, từ đó có thể thấy được đa phần các tuyến vận chuyển (khoảng 75%) là sớm so với dự kiến, còn lại là trễ khá ít. Tuy nhiên, ta thấy số lượng giá trị ngoại lai trễ ở đây rất nhiều.

Tổng hợp độ trễ theo từng tuyến:



Hình 5: Chi tiết độ trễ 4 tuyến vận chuyển theo thứ tự i1 → i4 từ trái sang phải

Nhìn chung thì cả 4 tuyến vận chuyển đều có tỉ lệ sớm cao, thuy nhiên tuyến i2 và tuyến i4 có số lượng và khoảng giá trị ngoại lai lớn so với khoảng phân vị, cũng là nguyên nhân gây ra các giá trị ngoại lai ở tổng hợp toàn quá trình vận chuyển.

Dựa vào sự phân tích trên, ta có thể đánh giá được công ty này hoạt động khá tốt do tỉ lệ sớm khá cao, tuy nhiên cần xem xét lại vấn đề có những tuyến quá trễ.

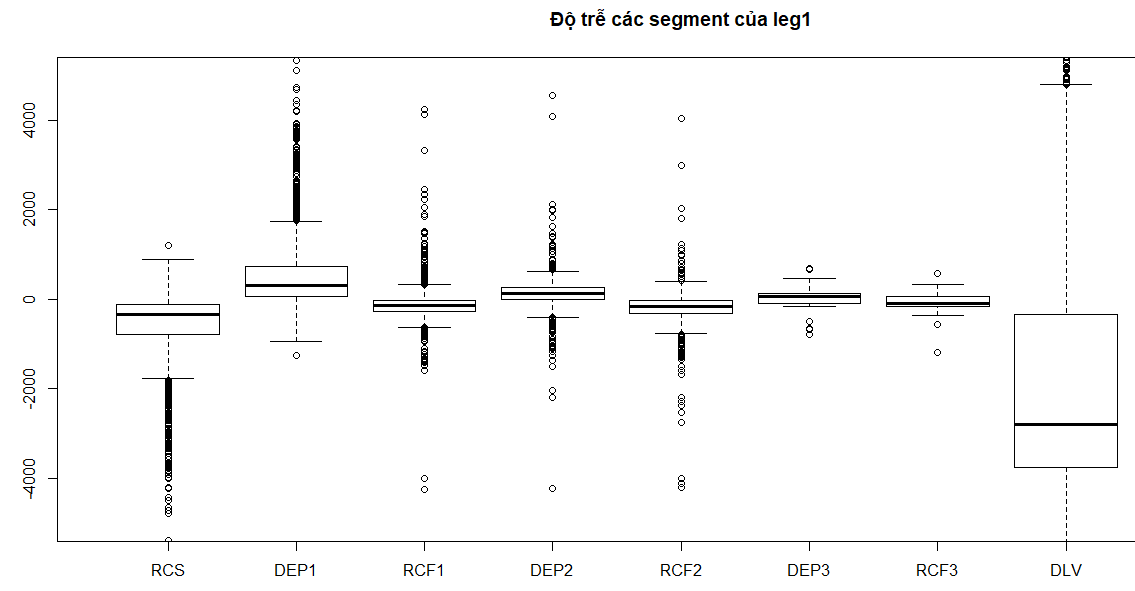
### Câu hỏi 2:

Từ câu hỏi 1, để làm rõ hơn chất lượng của công ty, ta thực hiện việc phân tích độ trễ của từng dịch vụ trong từng tuyến bay. Ta sẽ thực hiện đánh giá chất lượng từng hoạt động dịch vụ của công ty như thế nào?

Để trả lời được câu hỏi này, ta cũng sẽ dựa vào độ chêch lệch giữa thời gian thực hiện dịch vụ dự kiến và thời gian thực tế.

Trong bài báo cáo này, nhóm sẽ lấy một tuyến đại diện (i1) để trình bày kết quá, vì các tuyến còn lại cũng có kết quả giống như tuyến i1.

Ta sẽ thực hiện việc phân tích dịch vụ trên tuyến i1:



Hình 6: Độ trễ của từng dịch vụ trong tuyến đến i1.

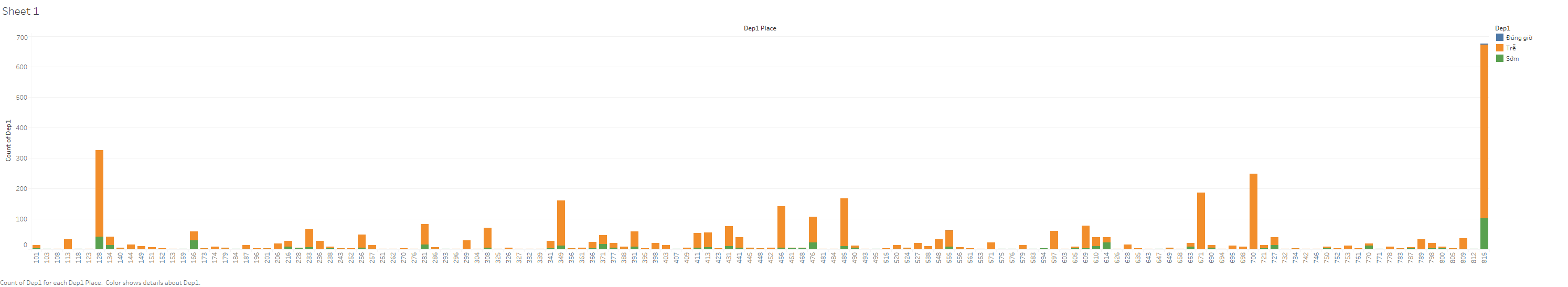
Phân tích hình 6, ta có thể thấy rằng dịch vụ DEP ( DEP1, DEP2, DEP3) là quá trình có hơn 75% số lượng tuyến bị trễ giờ so với dự kiến, còn các dịch vụ còn lại thì chưa đến 25% số lượng tuyến bị trễ.

Nhờ đó, ta có thể thấy rằng, dịch vụ DEP có chất lượng chưa được tốt. Nếu để cải thiện chất lượng của công ty, ta cần xem xét và nâng cao chất lượng của dịch vụ này.

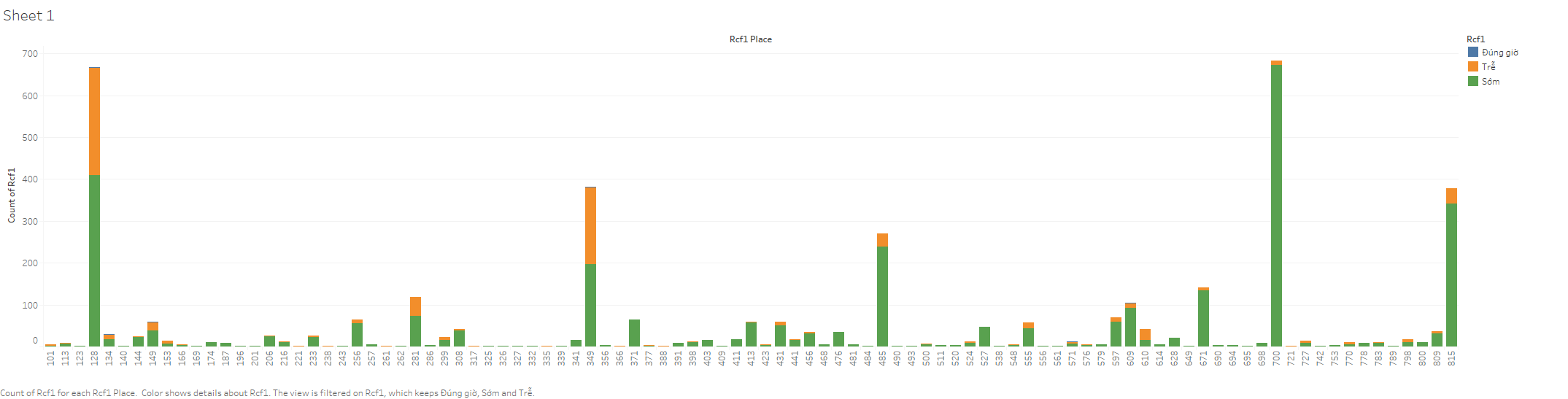
Câu hỏi 3:

Từ việc phân tích độ trễ của từng dịch vụ trong câu hỏi 2, ta có thể đặt thêm câu hỏi rằng: Nguyên nhân của việc trễ là do số lượng tuyến bay đến một địa điểm quá nhiều hay do tính chất hoạt động của dịch vụ đó.

Để có thể trả lời cho câu hỏi này, ta cần biết số lượng tuyến bị trễ giờ và sớm, cũng như đúng giờ của từng sân bay.



Hình 7: Số lượng tuyến bay i1 thực hiện dịch vụ DEP1 tại các sân bay



Hình 8: Số lượng tuyến bay i1 thực hiện dịch vụ RCF1 tại các sân bay

Hình 7 và hình 8 thể hiện số lượng tuyến bay thực hiện dịch vụ DEP1, RCF1 của tuyến i1. Tại mỗi sân bay sẽ thể hiện số lượng tuyến bay thực hiện dịch vụ sớm so với dự kiến ( thể hiện bằng màu xanh lá), số lượng tuyến bay thực hiện dịch vụ trễ ( thể hiện bằng màu cam) và số lượng tuyến thực hiện đúng giờ ( thể hiện bằng màu xanh dương).

Nhìn vào hình 7 ta thấy, việc thực hiện dịch vụ DEP (vận chuyển hàng hóa lên máy bay) đa số là trễ tại tất cả sân bay, phần màu cam chiếm đa số, dù là số lượng tuyến bay ở sẩn bay đó rất ít. Ngược lại, với hình 8, việc thực hiện dịch vụ RCF ( vận chuyển hàng hóa đến sân bay khác) đa số là sớm cho tất cả các sân bay, phần màu xanh lá chiếm đa số.

Từ đó, ta có thể thấy rằng, số lượng tuyến bay đến sân bay không ảnh hưởng đến chất lượng dịch vụ, mà sự chậm trễ của dịch vụ DEP là do tính chất hoạt động của dịch vụ đó.

# Cài đặt

Để trực quan hóa thông tin về các chuyến bay và chất lượng của từng dịch vụ, nhóm sử dụng dashboard. Dashboard là một cách đơn giản để tổ chức, quản lý nhiều biểu đồ trong cùng một bố cục và được tạo nên bởi cùng một dữ liệu cơ bản, giúp cho người dùng có cái nhìn tổng quát và nhanh chóng hơn.

Đây là hình của giao diện trực quan hóa của nhóm về bộ dữ liệu Cargo 2000.

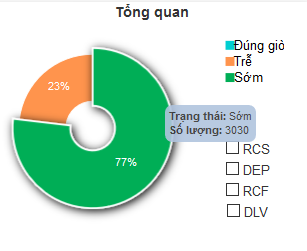
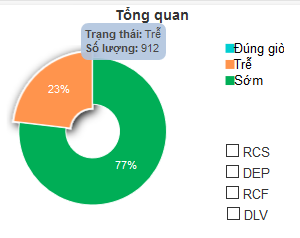


Hình 9: Giao diện trực quan hóa

Thông qua dashboard này , nhóm muốn thể hiện chất lượng tổng quan cũng như từng dịch vụ của công ty vận chuyển. Bên cạnh đó, dashboard còn thể hiện tần suất thời gian trễ / sớm của các dịch vụ cũng như độ trễ sớm tại các sân bay thực hiện vận chuyển hàng hóa.

Dashboard sử dụng một số các biểu đồ cơ bản để thể hiện các thông tin về bộ dữ liệu và cho phép người dùng tương tác để xem chi tiết hơn về thông tin chất lượng của từng dịch vụ. Dashboard dùng 3 màu chính: cam, xanh lá, xanh dương để thể hiện thông tin liên quan về độ trễ, sớm, đúng giờ của dịch vụ,

Đầu tiên, tổng quan về chất lượng dịch vụ thực hiện các chuyến bay vận chuyển hàng hóa của công ty được thể hiện dưới dạng biểu đồ donut. Biều đồ được chia thành 3 thành phần tương ứng với phần trăm các chuyến bay thực hiện trễ / sớm/ đúng giờ so với dự định. Khi đưa chuột vào từng thành phần trong biểu đồ này, ta sẽ thấy được số lượng chuyến bay tương ứng của mỗi phần

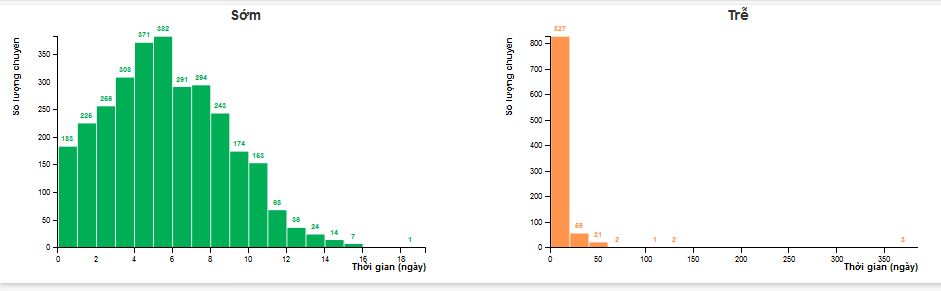
 

Hình 10: Số lượng chuyến bay tương ứng với từng phần trong biểu đồ tổng quan

Theo hình 10, ta thấy, số lượng chuyến bay thực hiện sớm hơn với dự định chiếm 77% tương ứng với số lượng là 3030, số lượng chuyến bay thực hiện trễ so với dự định là 23% tương ứng với số lượng là 912, và không có chuyến bay nào thực hiện đúng giờ.

Sau khi thể hiện tổng quan, dashboard biễu diễn thông tin chất lượng từng bốn dịch vụ chính của công ty bằng bốn biểu đồ phễu ( funnel chart ). Mỗi biều đồ phễu sẽ chứa thông tin về phần trăm số lần thực hiện dịch vụ trễ / sớm / đúng giờ so với tổng số lần thực hiện dịch vụ đó. Tương tự như biểu đồ tổng quan ở trên, khi đưa chuột vào từng phần của phếu, ta sẽ thấy được số lần thực hiện dịch vụ tương ứng.

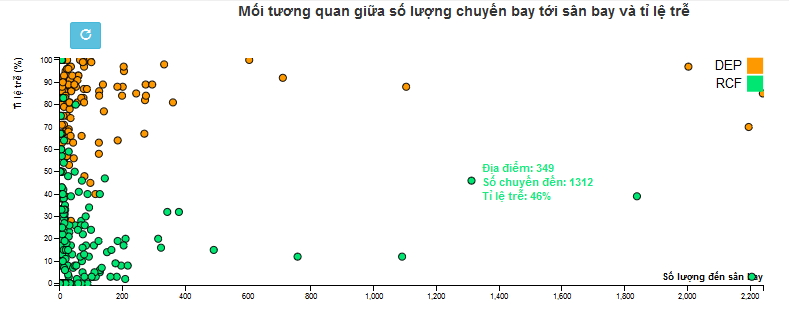
Để nắm rõ hơn về tình trạng trễ / sớm của các dịch vụ, dashboard thể hiện số lượng các chuyến bay thực hiện trễ / sớm theo khoảng thời gian bằng hai biểu đồ tần suất (histogram).



Hình 11: Biều đồ histogram thể hiện tần suất độ trễ / sớm của các tất cả chuyến bay

Theo hình 11, ta có thể thấy rằng, độ sớm của các chuyến bay chủ yếu vào khoảng 5-6 ngày. Bên cạnh đó, khi nhìn vào biểu đồ về độ trễ, ta thấy rằng có một số chuyến bay trễ hơn 350 ngày. Đây là những dữ liệu ngoại lai, công ty dựa vào đây có thể xem xét tại sao lại xảy ra tình trạng trễ nhiều như vậy.

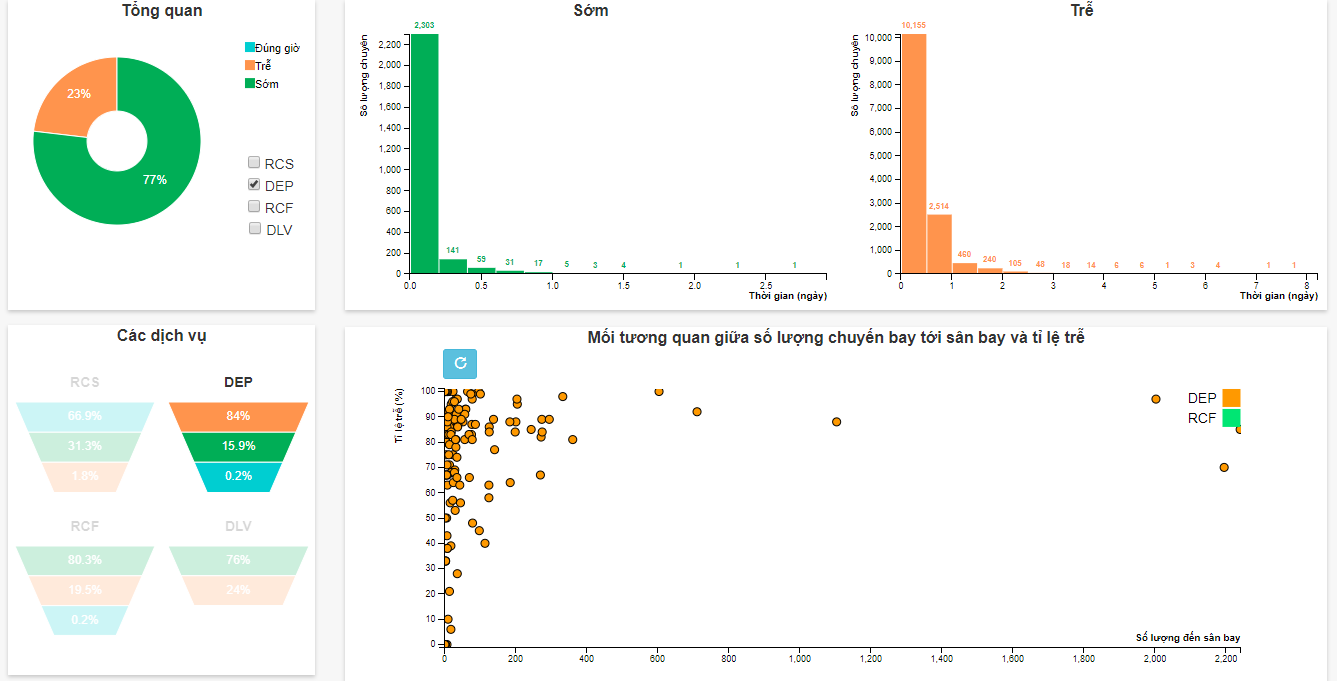
Để thể hiện thông tin đã phân tích được từ câu hỏi 3 ở mục V, nhóm dùng biều đồ phân tán ( scatter plot ) với mong muốn cho người dùng thấy được mối tương quan giữa phần trăm dịch vụ thực hiện trễ và số lượng tuyến bay thực hiện dịch vụ tại sân bay này. Vì trong bốn dịch vụ, chỉ có 2 dịch vụ DEP và RCF có dữ liệu về sân bay, nên biểu đồ này chỉ thể hiện dữ liệu cho hai dịch vụ này và được phân biệt bởi hai màu khác nhau.



Hình 12: Biều đồ phân tán thể hiện mối tương quan giữa số lần thực hiện dịch vụ và phần trăm dịch vụ thực hiện trễ tại các sẩn bay

Ở hình 12, ta thấy số lần thực hiện dịch vụ DEP tại các sân bay dù ít hay nhiều thì cũng phần trăm tuyến bay thực hiện dịch vụ DEP trễ tại sân bay đó đa số đều cao. Ngược lại, phần trăm tuyến bay thực hiện dịch vụ RCF trễ tại các sân bay đa số đều thấp. Nếu muốn xem rõ chi tiết về các điểm trên biểu đồ, ta có thể đưa chuột vào một điểm, biều đồ sẽ hiển thị thông tin cụ thể về mã sân bay ( địa điểm), số lượng tuyến đến và phần trăm số tuyến thực hiện dịch vụ trễ tại sân bay. Bên cạnh đó, người dùng có thể phóng to hoặc thu nhỏ một vùng của biều đồ bằng cách đưa chuột vào vùng đó và thực hiện zoom bằng chuột. Với tương tác này, người dùng có thể xem rõ hơn sự phân tán tại các vùng có mật độ phân tán cao và nắm thông tin rõ hơn. Từ biều đồ này, người dùng có thể nhận biết, số lượng tuyến bay đến một sân bay nhiều, không phải là nguyên nhân gây ra độ chậm trễ tại các dịch vụ. Từ đó, người dùng có thể tìm hiểu những nguyên nhân khác để cải thiện chất lượng dịch vụ cũng như chất lượng của công ty.

Ngoài ra, để thể hiện sự liên kết thông tin từ các biều đồ, dashboard cho phép người dùng thực hiện lọc dữ liệu theo dịch vụ bằng cách chọn dịch vụ từ các checkbox tương ứng. Khi một checkbox được chọn, thông tin về dịch vụ đó tại các biểu đồ phễu, biểu đồ tần suất, biểu đồ phân tán sẽ được hiển thị và thông tin các dịch vụ khác sẽ mờ đi. Nhờ vậy, người xem có thể nắm bắt rõ thông tin các từng dịch vụ qua các biểu đồ cùng một lúc.



Hình 13: Tương tác trên dashboard khi chọn xem thông tin về dịch vụ DEP

# Nhận xét và tự đánh giá

Giải pháp trực quan hóa của nhóm thể hiện được chất lượng nghiệp vụ của công ty, cũng như chất lượng của từng dịch vụ, phần nào thực hiện được mục đích ban đầu nhóm đề ra cho trực quan hóa. Tuy nhiên, giải pháp này chưa thể đưa ra nguyên nhân cụ thể của việc chậm trễ của các dịch vụ, mà chỉ có thể nêu lên yếu tố tuyến bay đến sân bay không ảnh hưởng đến độ chậm trễ này.

Tự đánh giá mức độ hoàn thành 8/10.

Tự đánh giá từng thành viên:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | MSSV | Họ tên thành viên | Vai trò | % công việc | Điểm |
| 1 | 1412503 | Nguyễn Thị Thanh Thảo | Thành viên | 25% | 8 |
| 2 | 1412542 | Nguyễn Hà Tiên | Nhóm trưởng | 25% | 8 |
| 3 | 1412543 | Nguyễn Thủy Tiên | Thư ký | 25% | 8 |
| 4 | 1412595 | Võ Thị Thanh Trúc | Thành viên | 25% | 8 |
| TỔNG | | | | 100% | 8/10 |