STAA57 Project Proposal

Group 14 - Aditya Sandeep Kulkarni, Alec Larsen, Md Wasim Zaman, Vishal Deb Sahoo

Link to the shared RStudio Cloud project that created this report:

https://rstudio.cloud/spaces/115177/project/2210076

Analysis Plan

With the data given and additional data on monthly air traffic, weather and seasons, the following questions will be adressed by our analysis:

How can scheduling be made the most efficient? We plan to suggest ways to enhance the quality and efficiency of scheduling training sessions by gauging patterns in the efficiency of instructors, duration between two sessions, the density of student enrollment every month, and student performance per season. The efficiency of instructors is assessed based on the average number of exercises students completed while being trained by each instructor. The density of student enrollment was estimated using the number of sessions that took place per month. The student performance was calculated based on the number of exercises completed by a student in a given time frame per session. All of these metrics indicate patterns that could help us increase the efficiency of scheduling sessions. Above results could also be used to optimize the recruitment of instructors with regards to the density of sessions per season.

What are the optimal conditions for student success? Our goal is to identify the conditions during training sessions that impact and, in turn, lead to student success. We consider a milestone like the first solo flight of the student as the factor expressing student training success. We account for a diverse array of metrics such as the number of exercises per hour of training, the weather and seasons during training sessions, the change in student performance for each instructor, the number of sessions required for the student to complete a milestone, most efficient training type, duration of particular exercises like dual flights, completion of a common set of exercises, comparison of students who didn't achieve particular milestones, etc.

In summary, our analysis plan is dependent on determining correlations between multiple important and controllable factors in the training process. We will do this by using data analysis tools such as graphs and linear regression to determine whether one factor of the training process impacts performance and if so, what the impact of that factor is.

Data

In answering these questions, we made the assumption in analyzing our data that all students begin their training at the same skill level. Additionally, we assume there were no unmeasurable external factors impacting the duration or efficiency of sessions. That is to say, unless a factor can be measured, it cannot be accounted for in our analysis.

To help determine when air traffic is the lowest, and thus flight accident risk is the lowest, we acquired air traffic data from the suggested supplementary data.
Air Traffic - https://open.canada.ca/data/en/dataset/b91772ed-edae-4fd4-8b80-a3e4c1d29976

We will utilise information in the following variables for observations of air traffic in Airports of Oshawa, Ontario for years 2015-2020:

- REF_DATE: Date of Reference provided in character "YYYY-MM" format which we will separate to two columns of type integer "Year" and "Month".

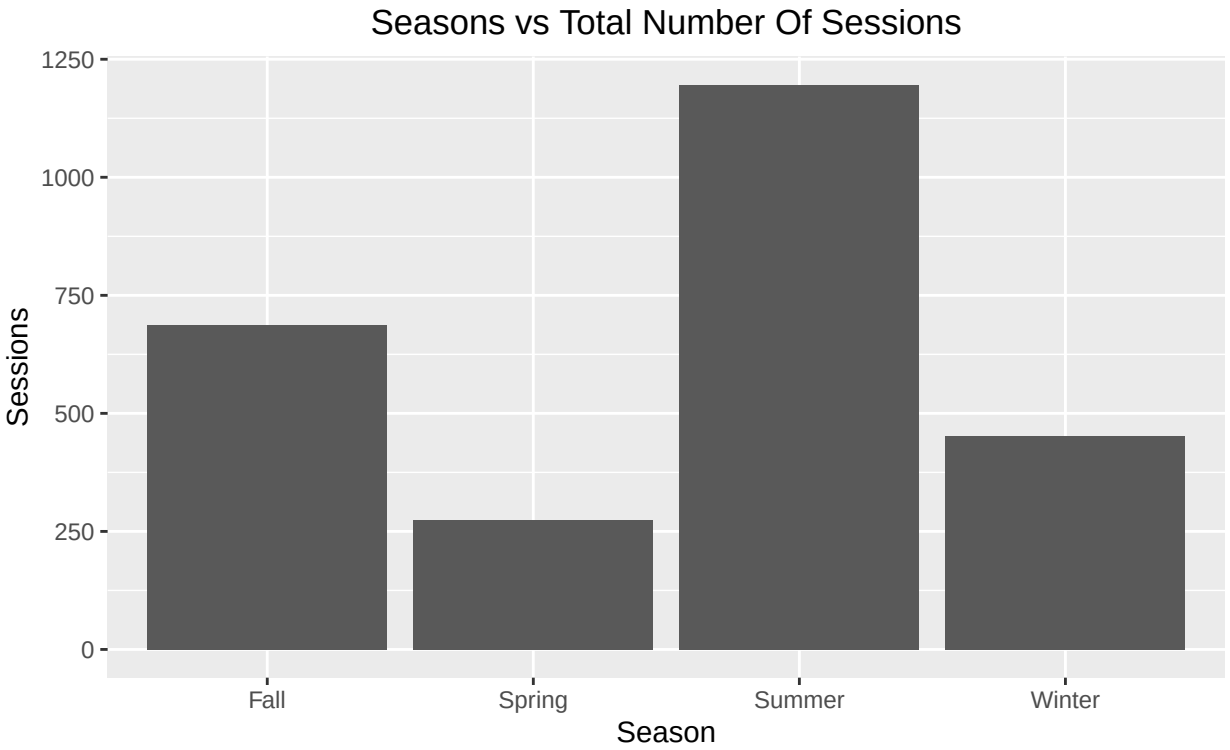
- `Civil and military movements`: Description of the air traffic.
- VALUE: The number of aiplanes.

To determine when flying is the safest (lowest wind, lowest precipitation), we used weather data from Government Of Canada's Website.
Weather data - https://climate.weather.gc.ca/historical_data/search_historic_data_e.html

We will utilise information in the following variables: Year, Month, Day, `Max Temp (°C)`, `Min Temp (°C)`, `Mean Temp (°C)`, `Total Precip (mm)`, `Spd of Max Gust (km/h)`

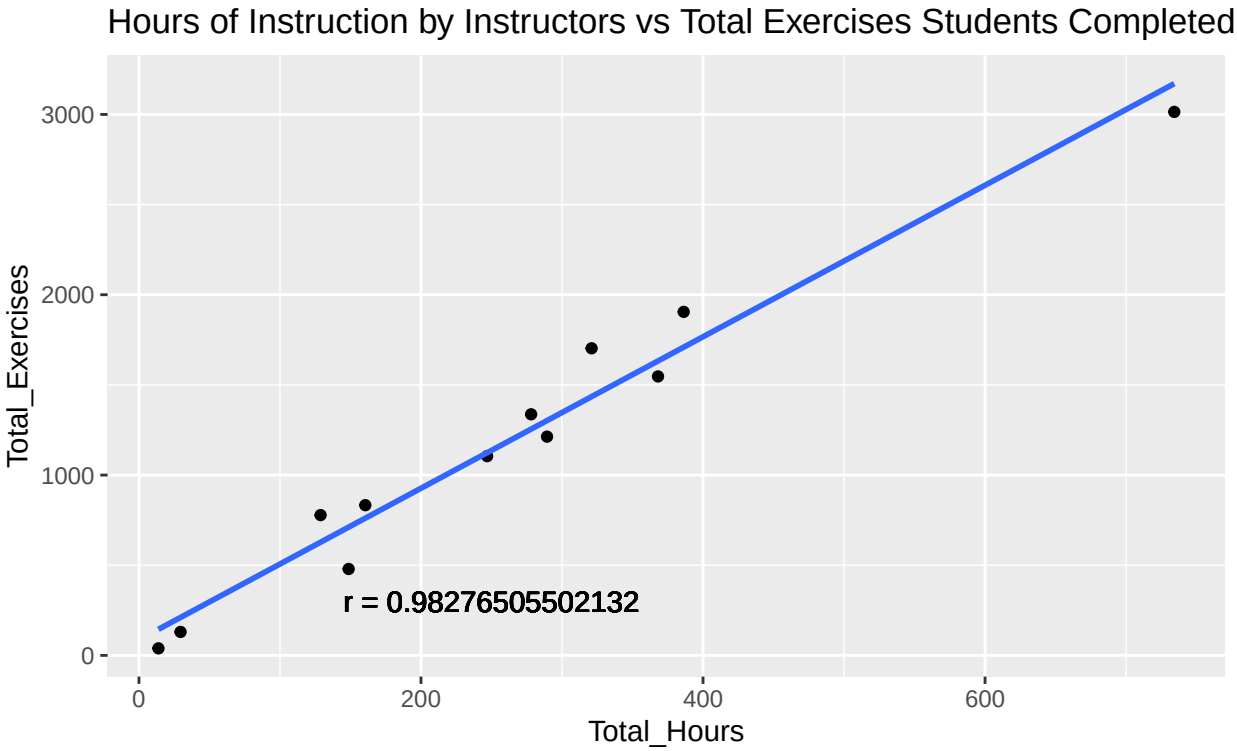
Preliminary Analysis

To suggest ways to improve scheduling we begin by understanding the current scheduling patterns. We graphed seasons against the total number of sessions conducted.



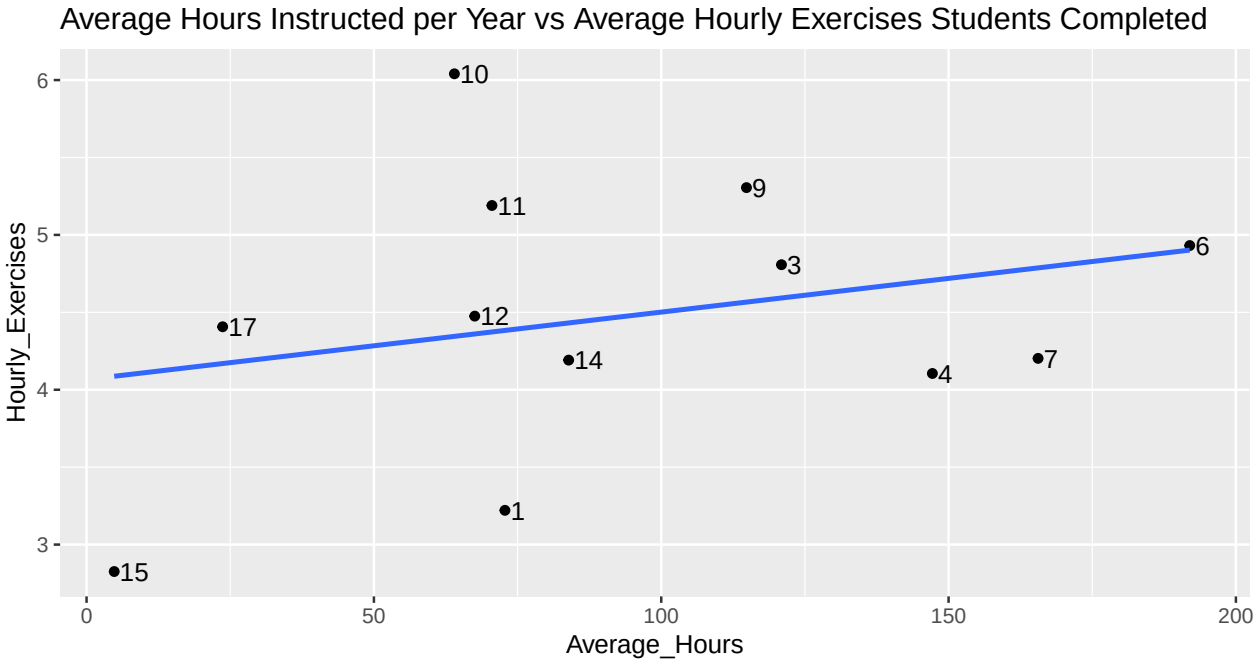
This graph clearly displays that the number of sessions conducted differed greatly from season to season with Summer being favoured the most. This suggests that either Summer presented better conditions for flying and hence for training sessions or it drew more customers or both.

To properly establish the most efficient method of training students, first we must establish a measure of efficiency. By graphing exercises completed in each individual session, we determined that on a session by session basis, the exercises completed cannot be related to the duration of the session. However, by graphing the total session hours completed against the total exercises completed under each instructors supervision, we found that there is a significant correlation between training time and the number of exercises completed in the long term.



Since both Total Hours and Total Exercises are roughly continuous, we can use calculate Pearson's Linear Correlation Coefficient. We found that r ~ 0.98, which shows strong positive correlation between the variables. Given this, we concluded that for analysis spanning several months or years, the number of exercises completed per hour is a reliable efficiency metric.

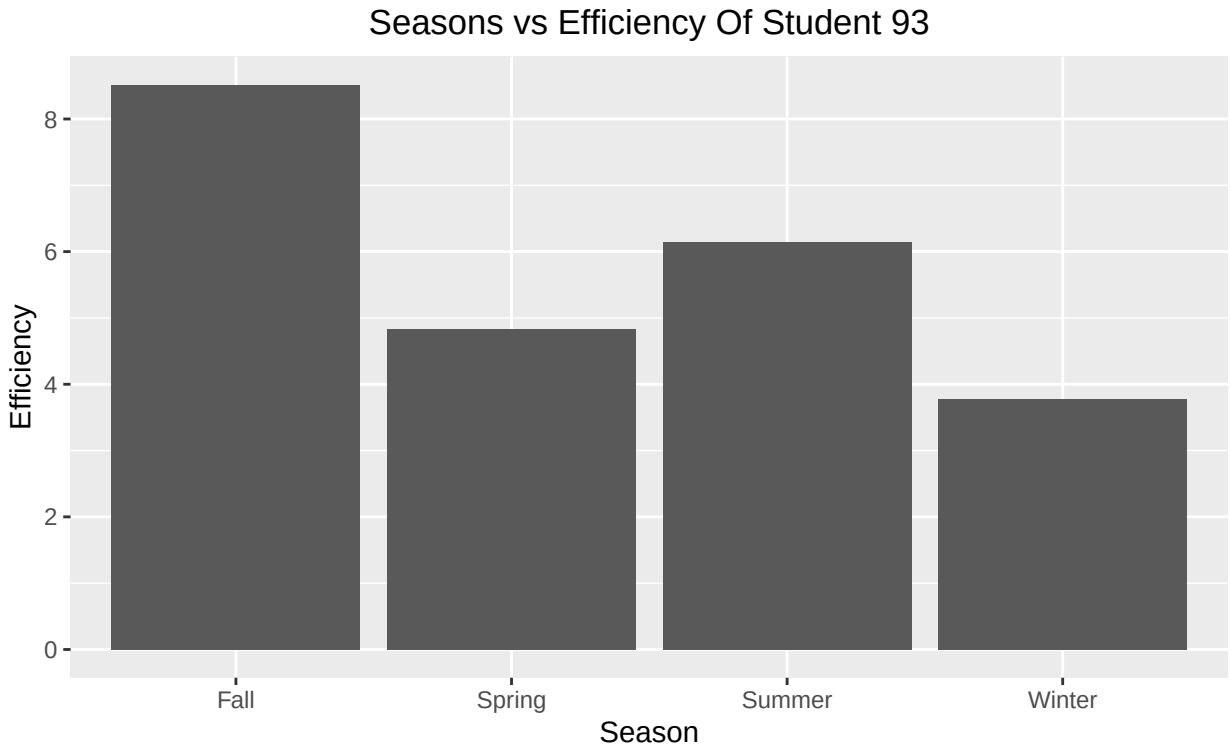
From this efficiency metric, we graphed the average hours each instructor worked per year against the average number of exercises students completed per hour.



This graph clearly shows no clear correlation between average yearly hours worked and the exercises completed per hour by students. This lead us to conclude that instructors do not necessarily get better the more

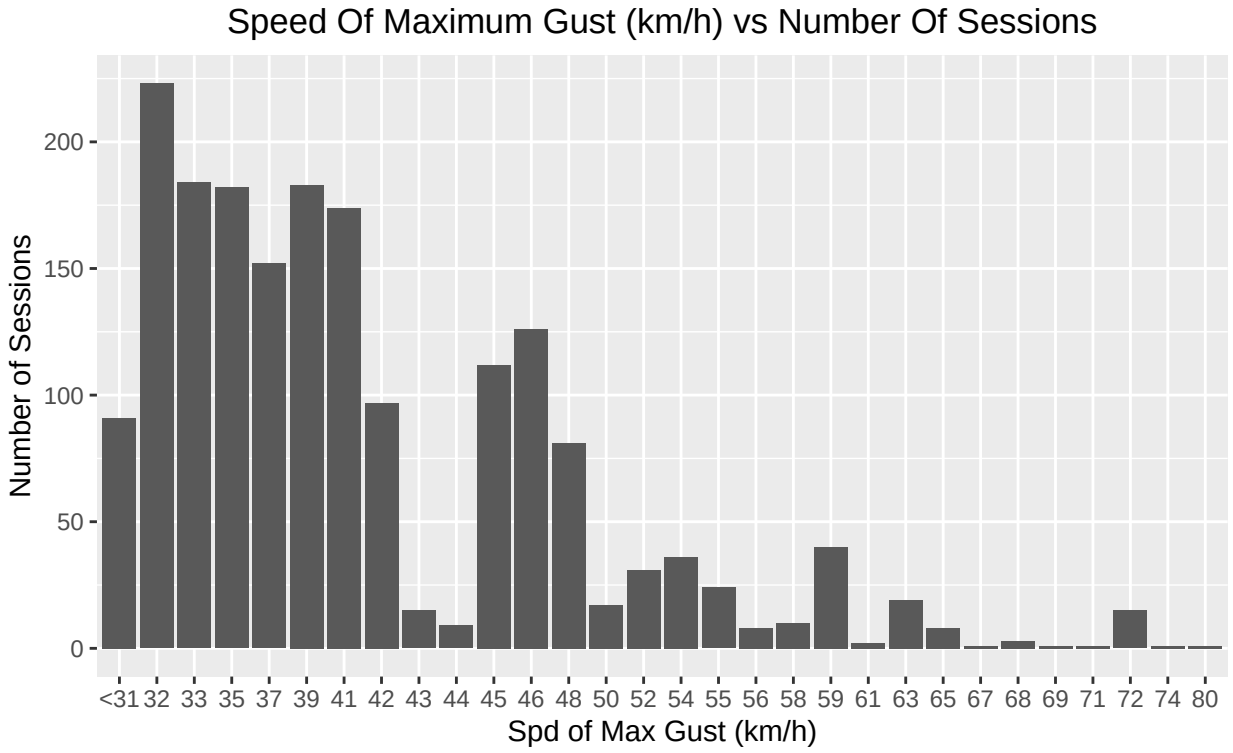
frequently they teach, but rather that some instructors are more effective than others. The line of best fit on this graph is not to show correlation but rather to plot the expected hours that each instructor should get based on their efficiency. Points above the line are, by our metric getting too few hours and instructors below the line are being scheduled for too many hours.

Given a significant difference between training activity between seasons and a fairly reliable efficiency metric, we graphed seasons against the efficiency of an arbitrary student (Student_ID = 93).



From this graph, we can infer that student's performance differed from season to season.

We graphed speed of maximum gust against the number of sessions.



This graph suggets that in general the num of sessions conducted decreases with increase in speed of maximum gust.

To establish that flying solo can be considered to be a metric for students progress we found the mean number of exercises completed by students that flew solo including the exercises performed during first solo flight(a) and also the mean number of exercises completed by students that never flew solo(b).

```
## [1] "mean (a)=  111.027777777778"
```

```
## [1] "mean (b)=  49.2289156626506"
```

We note that the mean number of exercises completed by students that flew solo is more than twice the mean number of exercises completed by students that never flew solo which suggests that there might be a certain level of experience or progress that a student must achieve before they fly solo.