STATISTICS SAVES THE DAY

SMU SINGAPORE MANAGEMENT UNIVERSITY

2nd MAJOR IN ANALYTICS
AN ANLY482 PROJECT

DARREN LIM FEI HONG
GOH GHEE GIN SHANE
LIM WEI YANG

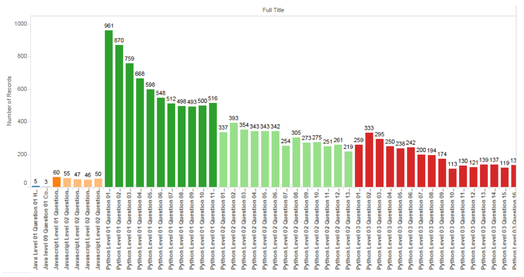
PROJECT INTRODUCTION

The Client – SingPath
SingPath is an educational website teaching Java, JavaScript, and Python. Students are given full autonomy to decide which questions to attempts and the order of attempts.

The Dataset – Firebase
With 1,544 unique users, and 223 unique questions, each row within the database is a unique attempt of each user on a question. There are 22,874 unique rows of attempts.

Language	Total Levels	Total Questions	Total Attempts
Java	2	2	8
Javascript	2	5	258
Python	10	216	22,608

Question Attempts by Language Level



Preliminary findings
The number of attempts per question generally falls as users move on to more difficult questions, as demonstrated in the graph above.

Many students code on weekends, specifically between Saturday evenings and Sunday morning.

PHASE 1: THE PREDICTIVE MODEL

Model 1: Predict via Median
For each question, what is the Median duration taken to solve it?

Model 2: Predict via Percentile
For each user, what is the average percentile they fall into when solving a question?

Comparison of Models

Model	SSE Score	Percentage of Correct Predictions
Median	49,127,814	44%
Percentile	41,340,317	56%

The Percentile Model appears to be more accurate!

Sum of Squared Errors (SSE)
The SSE score is directly correlated to the accuracy of the prediction model. It is the summation of the squared values of the difference between the predicted and actual values in a model.

MODEL LIMITATIONS

1. Only records <10 mins were used
2. Model is built on subset of data and is a poor representation of the dataset
3. Model poorly reflects the dirty nature of the data

PHASE 2: A REVISED MODEL

Examining the Source
Our team decided to dig into the source to discover why the data was so dirty and giving us no information.

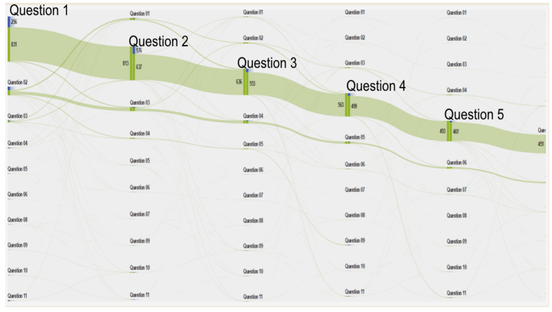
Discovered Limits
1. WEB PAGE
No enforced structure
No prompts for sequence
Many bugs on many pages

2. DATABASE
Lack of Dimensions

Current Sequence of Questions

Question Number	Question Title
5	Still More Variables
6	Variables
7	Another Variable
8	More Fun With Variables
9	Many Variables

Sankey Diagram of Questions



Students were meant to tackle questions according to their relative level of difficulty. However, instead the students were tackling questions in their presented sequence.

RECOMMENDATIONS

1. Add Dimensions to Database
This will allow for greater analysis of the data as more dimensions are added.

Column Name	Description
questionCompetencies	Competencies demonstrated by the user in order to correctly solve the question
userAge	The age of the user
userEducation	The education level of the user

2. Reorganise questions on SingPath
Arrange questions in order of competencies demonstrated

Proposed Sequence of Questions

Question Number	Question Title
6	Variables
7	Another Variable
5	Still More Variables
9	Many Variables
8	More Fun With Variables

3. Debug and Update SingPath.com
a) Fix cursor pointer in answer boxes
b) Redesign UI for intuitive use (i.e. Ensure buttons and boxes are functional)
c) Update questions to reflect competencies to be demonstrated
d) Add prompts for user to move on to the next question