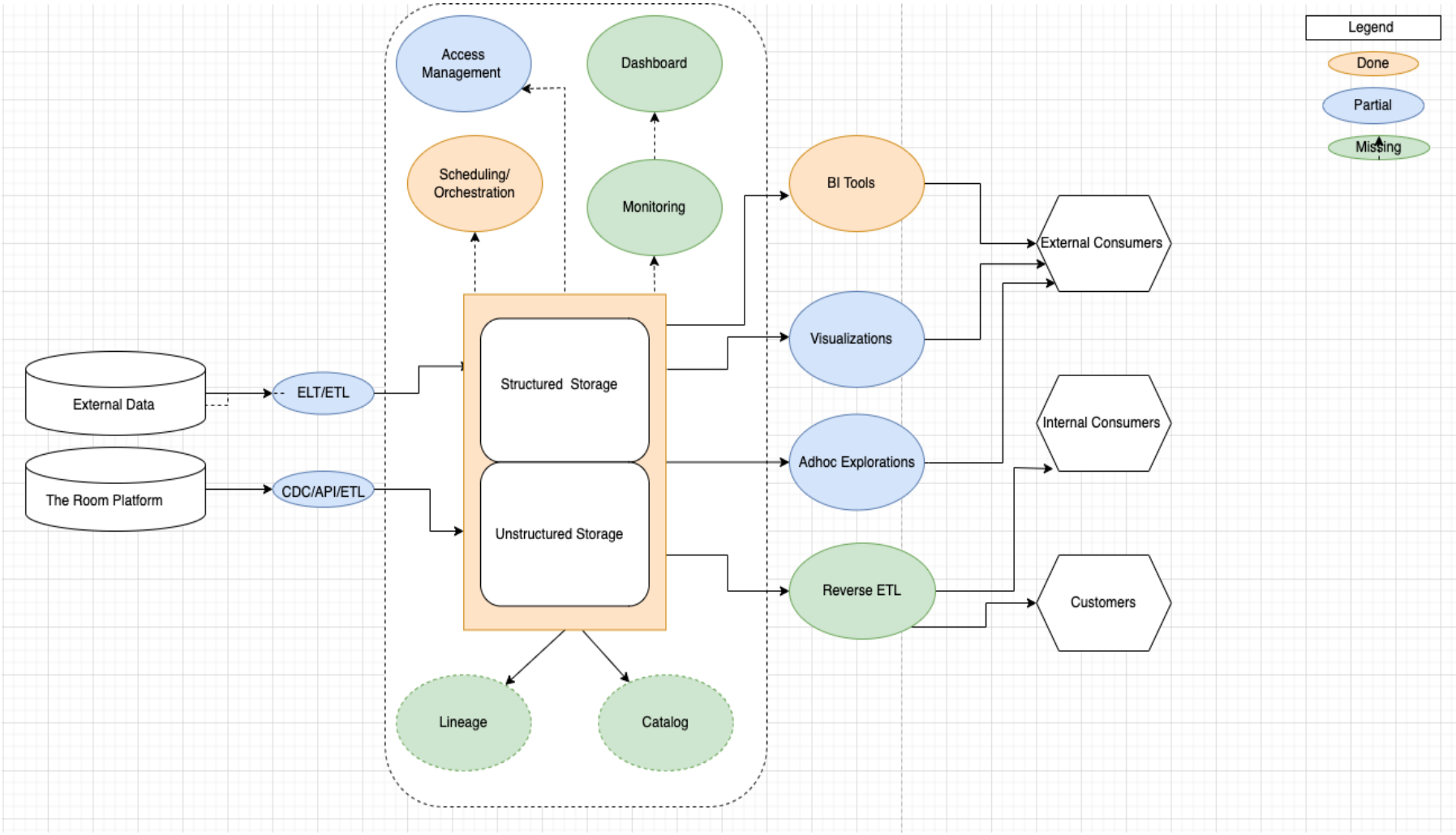
Data Platform Strategy

Mission: To create and manage a self-service data platform that is scalable, reliable, secure, frictionless and that people love to use.

Flow Chart



Classification of Tasks

The main focus here would be Data quality, Security and Monitoring in moving the data from user-facing infrastructure to the analytical end of the business.

Roles in the data platform

Data producers

Different services and applications. Create data contracts between Data Producers and Data Consumers on what the Data being produced should look like, what SLAs it should meet and the semantics of it.

Data contracts metadata will consist:

1. Schema of the Data being Produced.
2. Schema Version - Data Sources evolve, Producers have to ensure that it is possible to detect and react to schema changes. Consumers should be able to process Data with the old Schema.
3. SLA metadata - Quality: is it meant for Production use? How late can the data arrive? How many missing values could be expected for certain fields in a given time period?
4. Semantics - what entity does a given Data Point represent. Semantics, similar to schema, can evolve over time.
5. Lineage - Data Owners, Intended Consumers.

To Do:

1. *Outline purpose of data contract.*
2. *Outline implementation for data contract enforcement.*

Data extraction - Data Engineers

Extracting data from OLTP to OLAP.

Building monitoring infrastructure to enhance data reliability

Interfacing with data producers to establish contracts for downstream data expectation

Data modeling - Analytics Engineers

Provide clean, high-quality datasets so that different users within the company can work with them. This will entail transforming, testing, and documenting data.

Model datasets within business rules

Create utilities e.g., DBT macros and SQL functions within analytics for reusability and self service

Data/BI Analysis

Share data insights with business users, support self-service BI users, Build dashboards and reports.

Create and steward dashboards for company-wide metrics

Consultants for data related POCs

Help business stakeholder develop and validate hypothesis

Produce and interpret reports for external stakeholders

Provide feedback on data quality dimensions e.g., data availability and data freshness

Data Scientists

Carry out descriptive and predictive analysis, use machine learning techniques to improve the quality of data or product offerings and Communicate recommendations to other teams and senior staff.

Assigning of priorities to "Data Asks" from various teams

As a technical leader I will be focused on:

1. Creating a balanced portfolio of product work and technical debt (scale work and risk work)
1. Calibrating technical strategy with the nuance of company strategy.
2. Generating new ideas and make savvy trade-offs
3. Executing compounding strategies that ensure near-term technical initiatives deliver a big long term impact.

Roadmap

The road map will be influenced by the above 4 points. We will also consider what is **Done**, **partial** and **missing** on the architecture and classify it appropriately. Prioritization will also be done based on how much we impact the metrics that matter most for the customer/stakeholder and therefore for the business.

The metrics that matter most are:

1. Financial metrics:
    a. Revenue
    b. Profit margins
2. Customer metrics:
    a. Awareness
    b. Acquisition
    c. Activation
    d. Retention
    e. Referral
    f. Customer lifetime value
    g. NPS

Definition of metrics will be categorized into 2:

1. North star metric - what the business cares most which guides business decisions.
2. Supporting metrics - Metrics that improve the north star metric, to which we can tie project outcomes.

Example of a road map

To do	Product work	Scale work	Risk work
Catalog/ lineage (Missing) - We need to use dbt for our transformations as it provides an easy, version-controlled way of writing transformations using SQL. it also provides data quality checks natively. Thus, improving data engineering productivity, improving data quality and optimizing data governance.		✅	✅
Align on business Metrics and KPIS definitions (Partially Done) - Different definitions for same business metrics is a common occurrence. We need documentation to provide the different stakeholders with a single definition and perspective.	✅		
We need to incorporate office hours for the different stakeholders as most are incorporating data without context. For example, there was a spike in revenue caused by marketing running a campaign. Customer success stakeholders are not aware and think this comes from a customer training they carried out and might continue doing this.	✅		
We will incorporate metabase, a no-code solution that will empower different stakeholders to make data driven decisions thus freeing up data team to tackle bigger problems to establish and enhance data self-service.	✅		

Ad Hoc Requests

Ad Hoc requests will be looked at, requirements gathered, outcomes defined, prioritized internally and timelines communicated to the stakeholder.

Ad Hoc requests will be classified into three:

1. KPI definitions.
2. Data analysis.
3. Request for dashboards.

This will free up time for data practitioners to focus on the road map. Also, professional development through coaching and mentoring to different stakeholders as an investment for the team to tackle bigger problems, documentation to establish and enhance data self-service and alerts curation to better tackle data incidents and enhance data quality.

Course Analysis

Business case

John is the head of curriculum development at The Room. He wants to establish whether a course should be published at The Room or not. Currently he has no easy way of knowing how well a course will perform on the platform. He Invests a lot of resources in marketing campaigns to acquire students. John would want to know if a course will be paid for or not? how much? and how many students the course will have.

Business Objective

1. To perform descriptive and predictive analysis on the courses data.
2. To understand if price differs by content duration.
3. To investigate what level our subscribers are interested in.
4. To understand course level by content duration.
5. To Understand course level over the years.

Summary

1. Payment: Most courses in Udemy are paid. Web development and Business Finance are the most paid for courses.

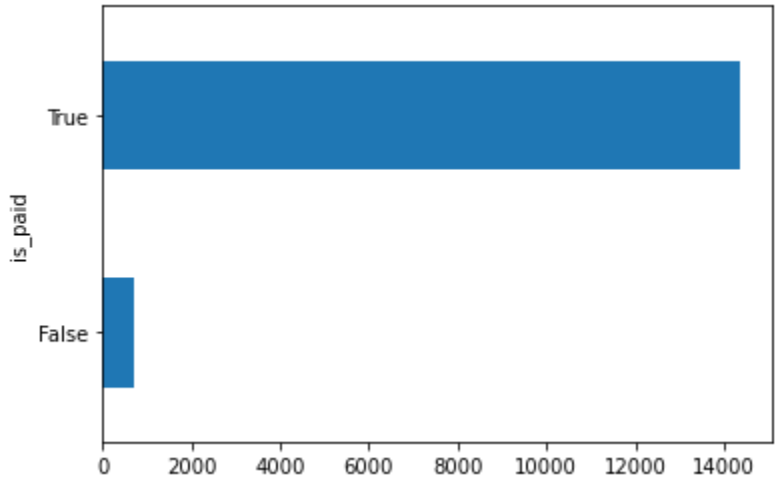
2. **Price and Duration:** We see that the content duration for Web Development with higher prices is high and for graphic design content duration is high with lower prices

3. **Number of subscribers**: over the years the number of subscribers has been decreasing.

4. **Level:** Almost 50% of courses available are of equal level for all learners. Then 35% are beginners level followed by 11% of intermediate level. Only 2% of entire course consist of expert level. Expert level courses have remained relatively the same over the years.

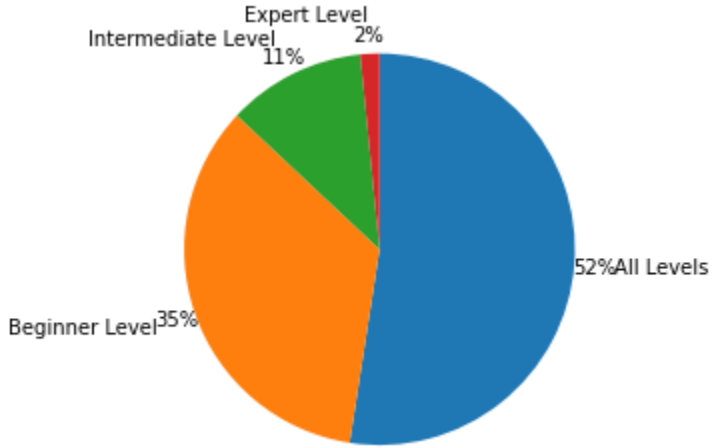
Observations

Content Duration and Payment



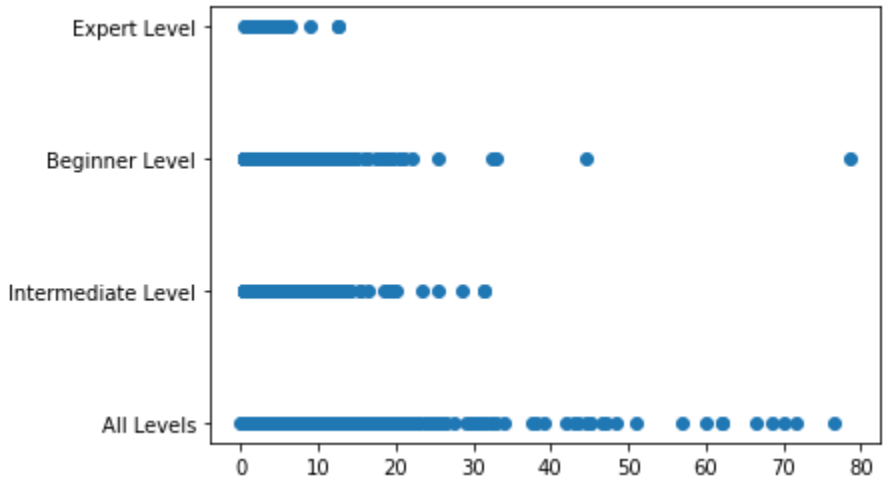
Course Level

Beginner level courses are at 35% followed by intermediate at 11% and Expert level at 2%

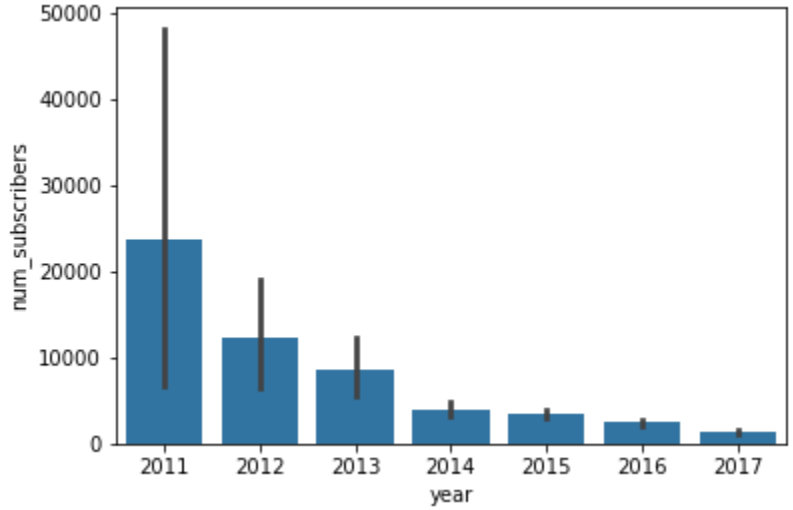


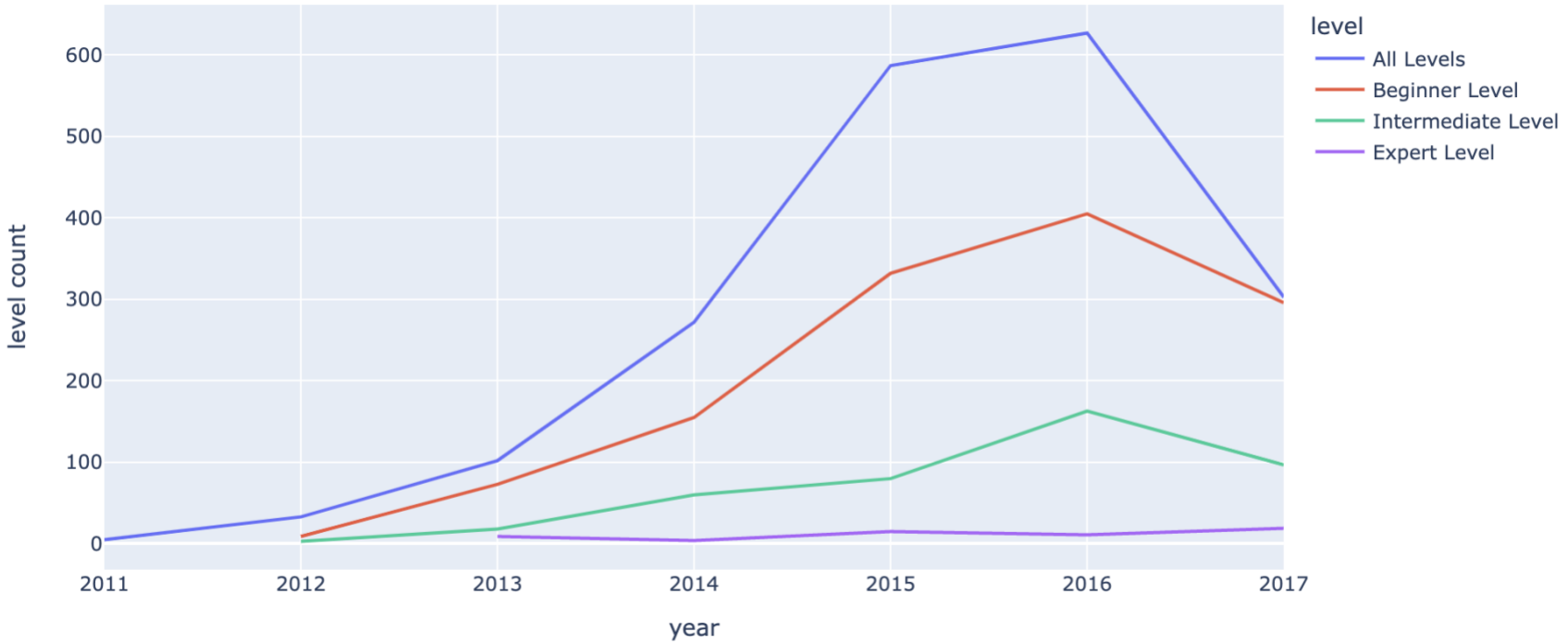
Course level by content duration

Expert courses are shorter in duration compared to beginner and intermediate level



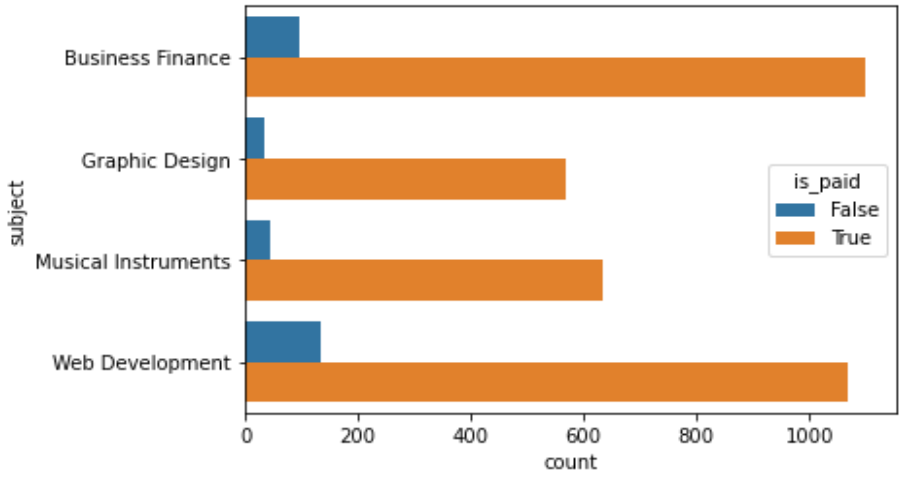
Number of subscribers over the years

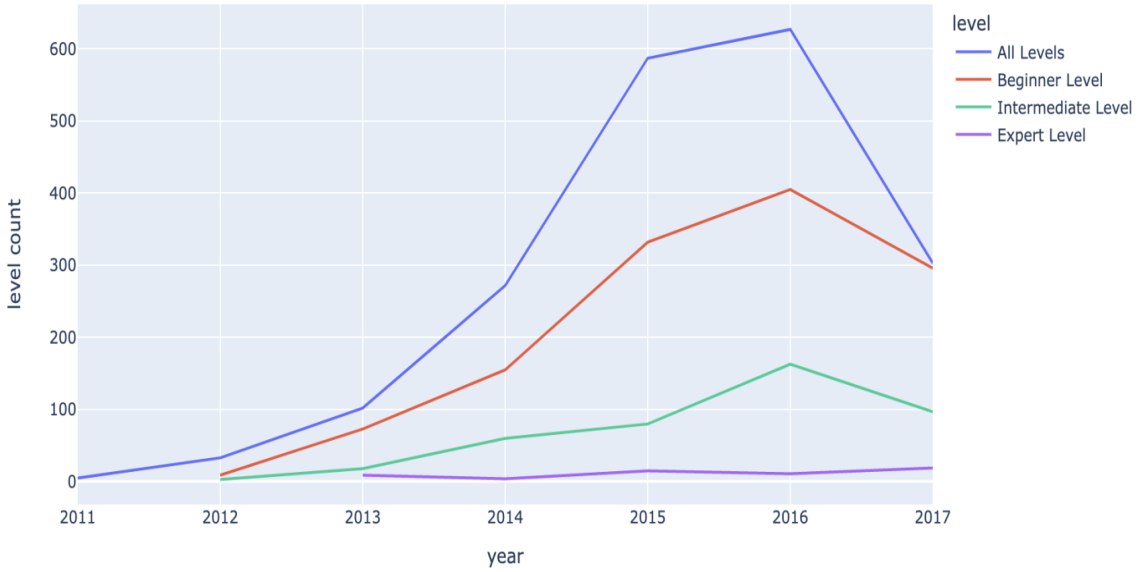



level
All Levels
Beginner Level
Intermediate Level
Expert Level
level count
600
500
400
300
200
100
0
2011
2012
2013
2014
2015
2016
2017
year

Subjects and payments

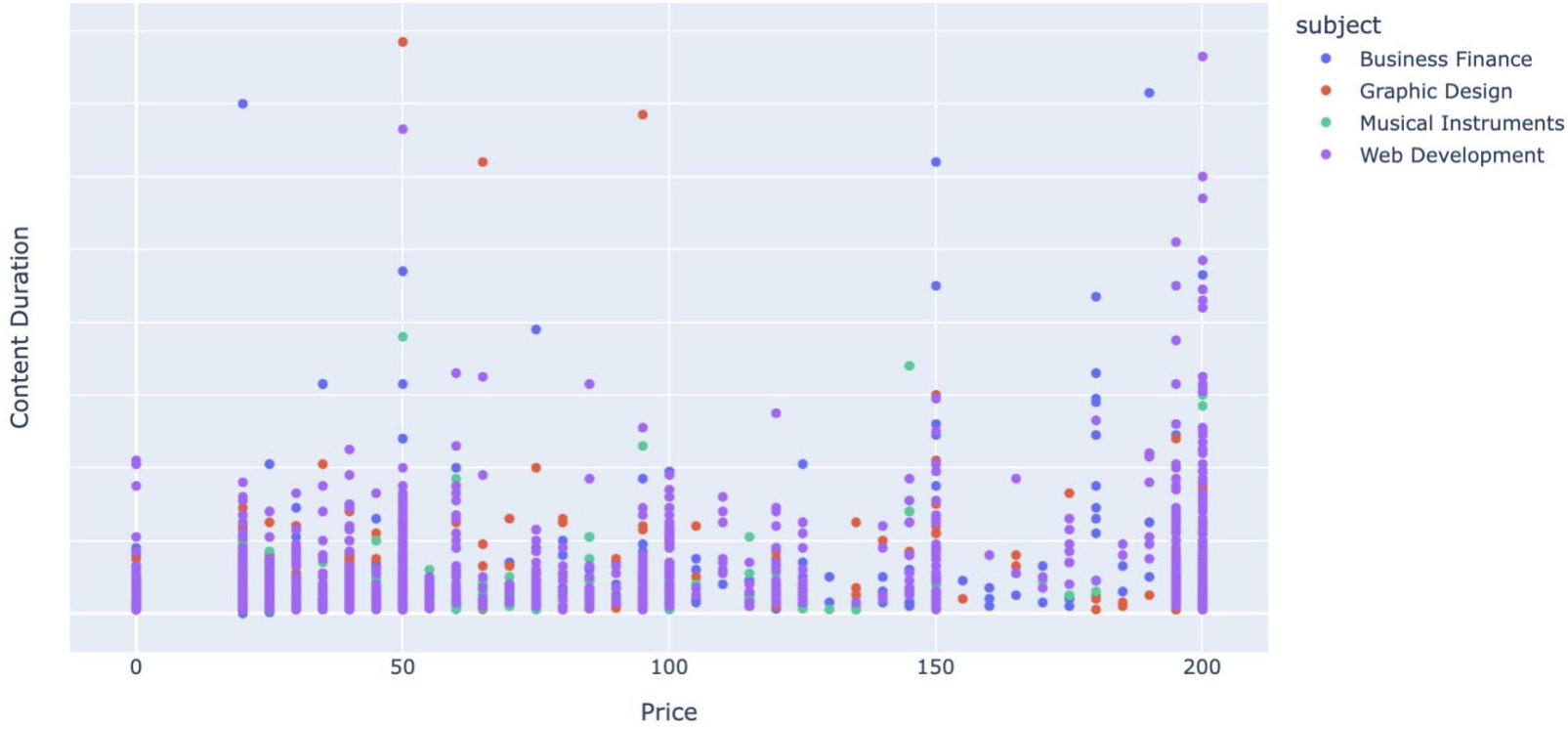
Web development and Business Finance are the most paid for courses.





Price and Content Duration

We see that the content duration for Web Development with higher prices is high and for graphic design content duration is high with lower prices



Recommendations

1. We should look into churn as from 2016.

Next Steps/Future Work

1. Build a recommendation engine as this will increase probability of a student taking a course hence increase in revenue.
2. Explore ratings of the courses.

Appendix

1. **Data**

This analysis uses data from 2011 to 2017

Machine Learning and Data Science Technical ROI Framework

Use case.

A Machine learning engineer is trying to measure the ROI of our just created model to predict if a course will be paid for or not. The goal would be to estimate cost savings per prediction.

Example

Assume we trained a decision tree classifier and below is our metrics report.

Accuracy 86%

Confusion matrix

[[8162 997]

[1560 7363]]

Formular

$$\hat{a} = (a - (1 - I) * e)$$

â - adjusted cost savings (profit per prediction)

a – expected cost

I – computed average accuracy (from training the model)

e- cost of listing all courses

to get the adjusted savings â we have to account for the ratio from the incorrectly predicted (1- accuracy) to the cost of making a mistake. The adjusted cost savings will give us the actual savings after removing the number of mistakes.

Assumptions

â = 3000 USD amount saved (Savings per prediction)

e – 8000 USD

I - 0.86

$$\hat{a} = (a - (1- I) * e)$$

$$\hat{a} = (3000 - (1-0.86) * 8000) = 188$$

We can expect to save 1880 per course assume we have 200 courses that amounts to 376,000 USD

Total adjusted cost savings would be 376, 0000

ROI Model and Decision-Making Framework for Course Resource allocation.

Use case

The Exec curriculum development wants to simulate the cost saving that will arise from publishing a course. As the data team we are tasked with coming up with a framework and providing a tool that he can use to determine if a course will be profitable or not.

Framework

Owner: Gerald Muriuki

Deliverable 1 - Framing

Problem Statement
John is the head of curriculum development at The Room. He wants to establish whether a course should be published at The Room or not. Currently, he has no easy way of knowing how well a course will perform on the platform both customer experience(good ratings) and financially. He Invests a lot of resources in marketing campaigns to acquire students. John would want to know if a course will be paid for or not? how much? and how many students the course will have.

Identify the critical decisions	To establish if a course will be paid for or not
Identify one or more possible outcome measures	Will have good ratings How much to charge
List the parameters	Price, level, content duration, Subject, Num of lectures, Time to publish
List the decision variables	Price he want to charge Time to publish Num of lectures
Indicate what you plan to leave out	The number of students on the platform and Number of simillar cources
Indicate what assumptions you will make	He cannot change the level He cannot change the subject He cannot change content duration
Indicate what is uncertain	Change of business model
Identify any intangibles	Which courses are currently in demand which courses are easier to get placement

Excel Scenario Simulator

The below excel wire frame simulates the calculation of expected return from a single course weighted over 12monhts.

The simulator can be used by the Exec in charge of Curriculum development.

Inputs							
Time on platform	46			Amount I will make after 12	$ 50,000		
Price	$ 145,000						
Num of Lectures	5						
Level	11						
content duration	30						
Subject	65						
Number of students							
Model							
Months	Price	Num of Lectures	Level	Num of Lectures	Subject	content duration	Number of s
1	$ 20	5.00	11	16	65	30	120
2	$ 20	5.00	11	16	65	30	100
3	$ 20	5.00	11	16	65	30	50
4	$ 20	5.00	11	16	65	30	40
5	$ 20	5.00	11	16	65	30	45
6	$ 20	5.00	11	16	65	30	48
7	$ 20	5.00	11	16	65	30	30
8	$ 20	5.00	11	16	65	30	30
9	$ 20	5.00	11	16	65	30	30
10	$ 20	5.00	11	16	65	30	30
11	$ 20	5.00	11	16	65	30	40
12	$ 20	5.00	11	16	65	30	30