AMDA Project 2016: Report

Promoting Entrepreneurship in India

*Our team has been appointed as an advisor to the PM to give suggestion on **'What are the important drivers to promote Business Entrepreneurship in India?'***

Problem 1

Regression

Problem 2

Conjoint Analysis & Clustering

Course structure guide & Capacity planning for IIMB

1) *Estimate **demand for elective courses** based on certain course attributes*
2) *Recommend an **optimal course structure** to faculty based on student preferences*

Group 1:

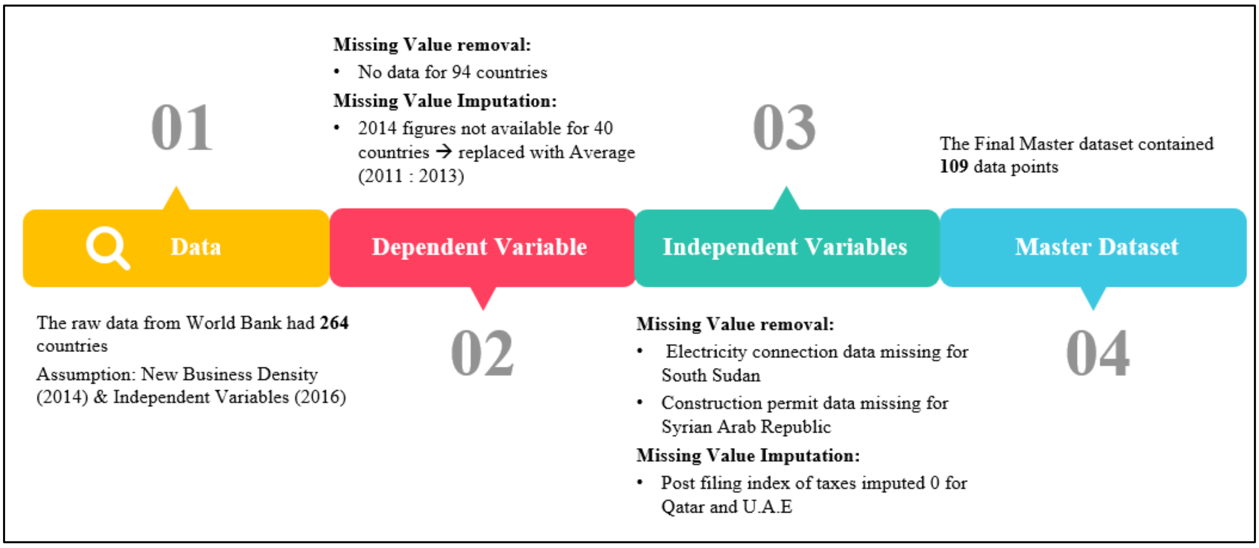
Abhishek 1511014 | Sagnik 1511049 | Sudhir 1511058 | Valentine 1511066

Identifying drivers to increase new business entry in India

1. **Premise:** Our team has been appointed as the advisor to the Prime Minister of India. And his agenda is to increase business entrepreneurship in India. There are countries who are performing reasonably well on this front, and countries performing terribly. We've been asked to figure out significant drivers that lead to the increase in the density of new businesses in any country and perform a sensitivity analysis to quantify the impact of recommendations. **Keywords:** Linear regression, correlation

2. **Methodology:** To solve the problem, we used a linear regression modeling approach to capture significant variables and their importance in influencing the growth of new businesses in any country. These variables are then used to perform sensitivity analysis and list tactical level recommendations.

3. **Data processing:** The data processing was the most crucial exercise, it involved collecting data from reliable sources to cleaning the data for any outliers/missing values to finally create a master dataset for subsequent analysis. Below is a snapshot of the data processing results:



01 Data

The raw data from World Bank had **264** countries

Assumption: New Business Density (2014) & Independent Variables (2016)

02 Dependent Variable

Missing Value removal:

- No data for 94 countries

Missing Value Imputation:

- 2014 figures not available for 40 countries → replaced with Average (2011 : 2013)

03 Independent Variables

Missing Value removal:

- Electricity connection data missing for South Sudan
- Construction permit data missing for Syrian Arab Republic

Missing Value Imputation:

- Post filing index of taxes imputed 0 for Qatar and U.A.E

04 Master Dataset

The Final Master dataset contained **109** data points

3.1. Data collection: For the purpose of this analysis, we required a dependent variable that measured the growth of new businesses in any country. The World Bank tracks the new business density for all the countries which we have used as a dependent variable for the analysis. The metric captures the number of new businesses per year per 1000 working population (age 15-64).

The data collection process for independent variables in the model was based on a hypothesis driven approach, where our hypothesis listed factors that lead to the growth of entrepreneurs in any country. These factors are broadly defined as ease of:

- Starting a business
- Obtaining legal construction permissions
- Getting electricity connection
- Registering property for a warehouse
- Access to credit facility
- Payment of taxes, and administrative tax burden
- Enforcing contracts

Based on these broad factors, we researched for specific indicators that define these factors. The World Bank tracks many indicators across all the countries, and is a reliable source for this study. We found 26 variables useful in defining the factors and have listed them as Exhibit 1 in the data analysis file.

3.2. Data mashing and cleaning: The World Bank tracks Doing Business parameters for 264 countries. The raw dataset is directly extracted from the World Bank website. The dataset contained New Business Density figures for a period of 2004 – 2014. The latest information about the dependent variable was available only till 2014. On the contrary, data about the independent variables was available only for June 2016. Hence there was a mismatch in the periods of dependent and independent variables. However, the New Business Density, being a national level figure, did not vary much year-on-year. Hence we've used 2014 figures for dependent variable, and 2016 for independent variable with the assumption that it will not impact outcomes significantly.

Missing value treatment: Below is the process followed for handling missing values in data:

Dependent variable	Independent variable
• Out of the 264 countries, 94 did not have the New Business Density figures for any year. These countries are not considered for the analysis. This set typically included small countries, countries defined as Offshore Financial Centers (OFCs) by IMF[1] and few exceptions such as China and Saudi Arabia. We believe that some of these countries might be hesitant to provide the data to World Bank. • Of the remaining 170 countries, 2014 New Business Density was not captured for 40 countries. We have imputed this missing value of 2014 by taking an average of the past 3-year data from 2011 to 2013.	• Data for the 26 variables mentioned in Exhibit 1 is traced from the World Bank website. We found that several countries did not have data for all the 26 variables. Hence we either removed these countries or imputed the missing values. Three such examples are mentioned below: ○ South Sudan has been removed from the dataset due to unavailability of Electricity Connection data. ○ Syrian Arab Republic has also been removed from the dataset due to unavailability of Construction Permit related data ○ The Post-filing index of taxes has been imputed as 0 for Qatar and U.A.E.

After the above adjustments, a final master dataset containing 109 countries has been used for the final post processing.

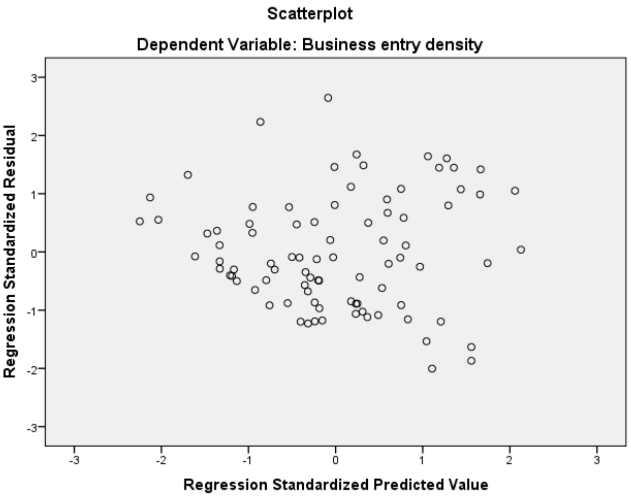
Outlier treatment: For all the variables, statistics like min, max, 99th percentile, 1 percentile was computed to detect outliers. However, the distribution of all variables were continuous, and it would have been inappropriate to remove countries falling outside the 1st / 99th percentile window.

[1] http://www.doingbusiness.org/data/exploretopics/entrepreneurship/methodology

4. **Analysis:** Once the master dataset was created, exploratory analysis on the data was conducted prior to modeling exercise to identify correlations between the variables.

4.1. Exploratory analysis: The correlations between the variables are present in Exhibit 2 in the data analysis file. The purpose of the exercise was to identify linkages between variable that needs to be dealt with using principle component analysis. However, no significant correlations were found between the independent variables.

4.2. Modeling: A linear regression based modeling was used to predict new business density using indicator variables discussed above. Given that variables had different dimensions, we standardized these variables using mean and standard deviation before using them in the model. Although some of the variables came significant, we faced a lot of challenges in correcting the model for heteroscedasticity and auto-correlation. A snapshot of standardized residual plot for one of the initial models is shown below, Exhibit 4 shows the detailed results.



To correct for heteroscedasticity and autocorrelation, we transformed the predictor variables before using them in the model. Following were some of the transformation we tested:

- Log of variables without standardization
- Square-root and cube-root of variables without standardization

However, all the transformations on independent variables resulted in heteroscedasticity. Hence we tried transforming the dependent variable using log, square-root, cube-root and fourth root.

The model using fourth root of dependent variable (business entry density) and standardized independent variables proved to be the best fit. The model was able to explain 57.6% variation in the dependent variable, and exhibited no heteroscedasticity and autocorrelation. The results of the final model are presented as part of Exhibit 3.

Regression equation:

$$Y^{1/4} = 1.12 + 0.06 * X_1 + 0.09 * X_2 - 0.07 * X_3 - 0.06 * X_4 - 0.05 * X_5 + 0.05 * X_6 + 0.05 * X_7 + Error$$

Where Y is Business entry density and standardized X_i's are:

X_1 : Property registration - Quality of the land administration index (0-30)

X_2 : Contract enforcement - Quality of judicial processes index (0-18)

X_3 : Electricity connection - Cost (% of income per capita)

X_4 : Ease of starting a business - Number of Procedures

X_5 : Tax system - Total tax rate (% of profit)

X_6 : Credit facility - Credit registry coverage (% of adults)

X_7 : Tax system - Post filing index (0-100)

4.3. Inferences from the model: Below are the inferences obtained from the model:

Factor - Indicator	Importance	Comments
Contract enforcement - Quality of judicial processes index (0-18)	0.28	A strong judicial process indicates smoothness in commercial dispute resolution, and impacts positively for new businesses
Electricity connection - Cost (% of income per capita)	-0.22	High cost of obtaining electricity connection hinders setting up of new businesses
Ease of starting a business - Number of Procedures	-0.19	High number of procedures (formalities) acts as deterrent for new businesses
Property registration - Quality of the land administration index (0-30)	0.17	A good land administration impacts positively for new businesses by helping them acquire land quickly
Tax system - Post filing index (0-100)	0.15	Hassle-free post tax filing process (refunds etc.) has a positive impact on the growth of new businesses
Tax system - Total tax rate (% of profit)	-0.15	Higher tax rate acts as a barrier for new businesses
Credit facility - Credit registry coverage (% of adults)	0.14	High coverage of credit registry facilitates credit lending, thereby supporting new businesses to raise funds for start-up

4.4. Sensitivity analysis: The purpose of the analysis is to quantify the impact on business entry density by changing the underlying indicator variables. Let's assume Y is a function of only one variable X_1. Then:

$$Y^{1/4} = \beta_1 * \frac{(X_1 - \mu_1)}{\sigma_1} + \ Constant$$

$$\frac{\Delta Y}{Y} = 4 * \frac{\beta_1}{\sigma_1} * \ \Delta X_1 * \ Y^{-1/4}$$

Factor - Indicator	Value (India)	India's rank –parameter (out of 109)	Change	Impact on new business density
Contract enforcement - Quality of judicial processes index (0-18)	9	45	Increasing the index by 1 pt	21%
Electricity connection - Cost (% of income per capita)	133.2	43	Reducing electricity connection cost by 50%	2%
Ease of starting a business - Number of Procedures	12.9	104	Reducing number of procedures by 1	14%
Property registration - Quality of the land administration index (0-30)	7	94	Increasing the index by 1 pt	5%
Tax system - Post filing index (0-100)	4.3	105	Increasing the index by 1 pt	1%
Tax system - Total tax rate (% of profit)	60.6	100	Reducing tax rate by 10 percentage points	22%
Credit facility - Credit registry coverage (% of adults)	0	104	Increasing credit registry coverage by 1 percentage point	1%

5. Recommendations:

Based on the sensitivity analysis, following are the recommendations to increase the number of new businesses in India:

– **Tax system: Tax exemption for initial years of set-up, with easier refund process.**
India ranks #100 out 109 in the tax rate. A tax exemption on initial years is likely to promote the growth of entrepreneurs in India.

– **Start-up formalities: Reducing the number of new business registration formalities**
India has too many formalities that entrepreneurs need to fulfil to register their businesses. The mean number of procedures across different countries is 6.43, which is far low than what we have in India (12.9). The registration process needs to be fast-tracked with lesser amount of documentation to promote new businesses.

– **Property registration: Need to reduce administrative hassles in land acquisition**
India ranks #94 in quality of the land administration index. Businesses who need to setup their warehouses get stuck in the administrative process. For e.g., mutation process, i.e., conversion of agricultural/residential land to commercial, takes a lot of time. Administrative hassle can be reduced by clearly defining the mutation process and reducing the number of administrative approvals required (single point clearance).

– **Contract enforcement: Fast-track commercial dispute resolution required**
India's quality of judicial process index ranks #45 in the world. Most of the new businesses fear of getting stuck in the courts with their disputes due to high number of pending cases. Fast-tracking resolution of commercial disputes can significantly impact growth of new businesses.

Identify optimal course structure for IIMB students across various streams

1. **Premise:** Our team is part of the academic council at IIM Bangalore. One of our responsibilities include planning for elective courses that students bid for in Terms 3,4,5 and 6. It is known that students are made privy to course feedback from previous years before they are asked to bid for electives. A certain set of parameters for courses are considered by students before bidding. For instance, some students enjoy quantitative courses and are willing to take any and all that are offered, while some students are great at participating in classroom discussions and hence prefer courses with high CP components. We look at two specific objectives, one, to estimate demand for a course given these parameters (described later), and two, to recommend optimal course structure based on student preferences to faculty of various departments.

2. **Survey Design:** Based on certain attributes for courses that are felt to significantly impact course bidding decisions (acquired from experience of being in academic council), we asked students to rank a select few sample courses which vary widely in values for these attributes. We communicated these attributes to the students as part of the survey and asked them to rank these courses in order of preference. These attributes are described in the following table:

Attribute	Description and Values
Course Pedagogy	What tools the faculty uses to teach in class over the term. *Theory Based or Case/Project based.*
Quantitative Level	The level of quant in the course (some people hate quant!) *Low, High*
Work Load	Whether components are evenly distributed across term or heavy MT/ET *Even, Heavy MT/ET*
Class Participation	Dictates the importance of CP (and the level of DCP!) *Low, High*
Ownership of Group Formation	Who makes groups for project/case work in the course? *Student, PGP Office/ Faculty*

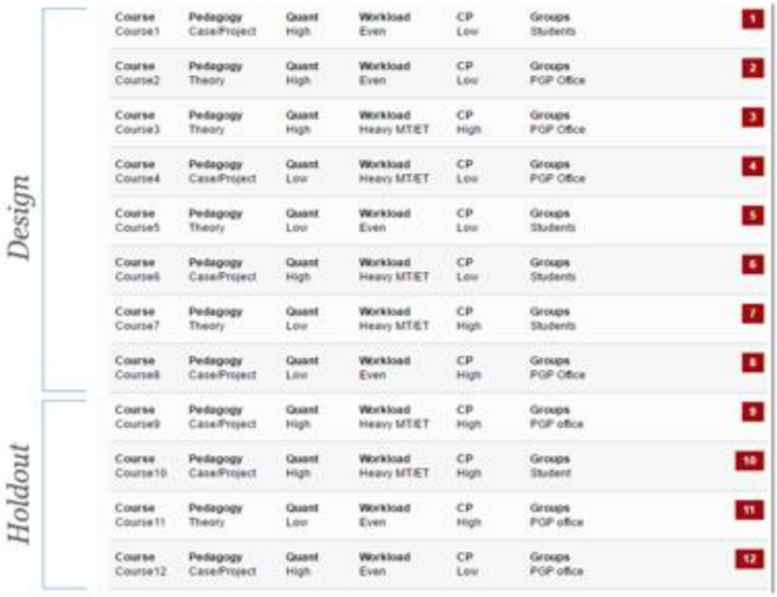
We also asked students to include information like their preferred stream of courses. This would help us advise faculty for each department as to what students, who prefer that stream, want the course structure to be like. Basic demographic information was obtained too; though there was no intention of using this, this was more to be sure that our data was drawn from a fairly homogenous population. Based on the levels of these attributes we used SPSS for an orthogonal design. The output was 12 cards in total, 8 design and 4 holdouts.

3. **Data Collection:** The next step was to get people to fill this survey. This turned out to be significantly harder than we thought,
   - As described above, there were 12 cards that we needed people to rank. **UI design** for this became a challenge. For one, the idea was to show all 12 cards on the same screen along with full attribute information for each card to facilitate comparison. It was tough to get this to work on desktop web browsers, also because the survey application (Qualtrics[2]) did not support such design out of the box. We ended up spending quite some time writing raw HTML, to make this possible (see Exhibit 5 for sample). Another issue was with mobile screens, which is generally a favoured method for people to fill surveys on, this was turning out to be intractable.

[2] https://iimb.au1.qualtrics.com/ControlPanel/

– Because of the UI issues, and because of how complex the survey was in its essence, very few responses came in in the first few days. This was alarming to us because we had set our sights on at least 90-100 responses on which to perform our analysis. We started personally visiting people's rooms, cashing in on favours from other people, asking PGP1s we'd helped in some way or the other to take the survey. These efforts paid off eventually and we had our ~100 responses within a day.

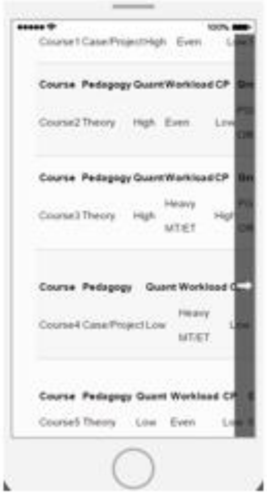
Cards obtained through orthogonal design



Course	Pedagogy	Quant	Workload	CP	Groups	
Course1	Case/Project	High	Even	Low	Students	1
Course2	Theory	High	Even	Low	PGP Office	2
Course3	Theory	High	Heavy MT/ET	High	PGP Office	3
Course4	Case/Project	Low	Heavy MT/ET	Low	PGP Office	4
Course5	Theory	Low	Even	Low	Students	5
Course6	Case/Project	High	Heavy MT/ET	Low	Students	6
Course7	Theory	Low	Heavy MT/ET	High	Students	7
Course8	Case/Project	Low	Even	High	PGP Office	8
Course9	Case/Project	High	Heavy MT/ET	High	PGP office	9
Course10	Case/Project	High	Heavy MT/ET	High	Student	10
Course11	Theory	Low	Even	High	PGP office	11
Course12	Case/Project	High	Even	Low	PGP office	12

*Design Cards – **8**, Holdout Cards – **4***

Issues while viewing survey on mobile


Course1 Case/Project High Even Low
Course Pedagogy Quant Workload CP
Course2 Theory High Even Low
Course Pedagogy Quant Workload CP
Course3 Theory High Heavy MT/ET High
Course Pedagogy Quant Workload
Course4 Case/Project Low Heavy MT/ET Low
Course Pedagogy Quant Workload CP
Course5 Theory Low Even Low

The above limitations betray a significant drawback of conjoint analysis in our opinion. Survey design needs to be smart, psychometric tests for instance display questions one after the other after which a rank is probably computed. Asking users to rank all cards at once is fraught with difficulties and should be avoided if possible. Some sort of incentive, financial or otherwise, could be a good way to coax users into filling these surveys as genuinely as possible.

4. **Data processing:** 99 responses were recorded in all. Exhibit 6 shows the distribution of data across various categories. Because the survey was complex and fairly cumbersome to take, there were some issues where people had not changed enough rankings and had submitted the default list as is. This was expected, we had planned for this and had devised a strategy wherein if less than 4 original card rankings were modified we would ignore the corresponding survey entry as a whole. After this exercise we were left with 87 unique rank orderings.

5. **Analysis:** The next step was to proceed with conjoint analysis followed by clustering. The objective was to identify clusters of students who had a preference for certain types of course outlines and to recommend faculty preferred course structures for their department.

5.1. Conjoint analysis: We wrote a code in SPSS to perform a full-profile conjoint analysis. This code is attached in Exhibit 7. The inputs were the ranked responses (with the 'SEQUENCE' keyword which meant that Rank1 was the 1st ranked card and not the rank of the 1st card) and orthogonal plan was used to design the cards. As output we obtained the utility tables and Kendall's Tau significance tables for each response. The aggregate correlations table, importance levels and utility summary are given in Exhibit 8.

Conjoint analysis equation: Based on the average utilities, below is the conjoint analysis regression equation, however the equation will be different for different variables:

$$Y = 4.5 - 0.046 * X_1 + 0.328 * X_2 - 0.880 * X_3 - 0.46 * X_4 - 0.166 * X_5 + Error$$

Where Y is rating of across different cards and X_i's are dummy variables:

X_1 : 1 when quant is high; 0 when quant is low

X_2 : 1 when pedagogy is case/project based; 0 when pedagogy is theory based

X_3 : 1 when workload is heavy midterm/endterm; 0 when workload is even across the term

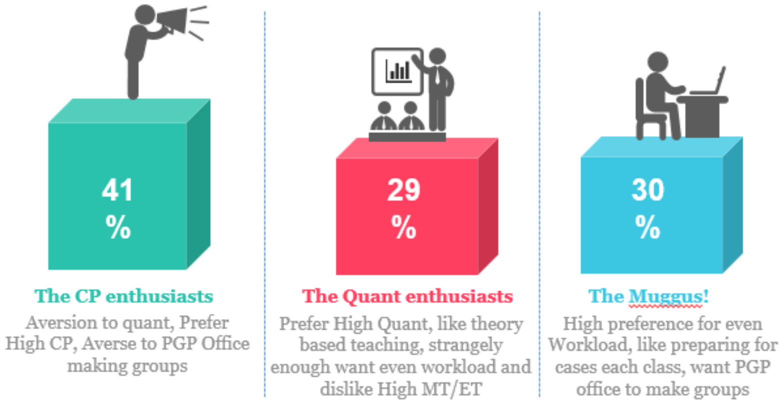
X_4 : 1 when CP is high; 0 for low CP

X_5 : 1 when PGP Office/ Faculty forms project groups; 0 when students can form their groups

A quick summary of the results,

- Significant Kendall's tau for the utility values of all attributes across all students
- Survey-responses are consistent based on Kendall's tau for holdout cases
- Student look for Quant, Workload and CP in respective order before taking the course; Pedagogy (whether course is theory or case based) is the least important attribute.
- The average utility values do not depict the whole picture. It made sense to follow this up by drilling one level deeper, i.e. performing clustering analysis.

5.2. Clustering: The next step was to perform clustering, as the output of conjoint analysis readily seems to suggest. We first started with hierarchical clustering to get a better sense of how many clusters to look for. The dendrogram obtained is given as Exhibit 9. It was clear from the dendrogram that it made sense to look for 2, 3, or 4 clusters. We then tried the K-Means clustering approach with the objective of finding 2, 3 or 4 clusters based on the Euclidean distance method. The 3 cluster quest gave us the best results, and 3 disparate clusters were identified. Information about these clusters (definitions, and average utility values across different attribute levels) is given below.


41 %
The CP enthusiasts
Aversion to quant, Prefer High CP, Averse to PGP Office making groups
29 %
The Quant enthusiasts
Prefer High Quant, like theory based teaching, strangely enough want even workload and dislike High MT/ET
30 %
The Muggus!
High preference for even Workload, like preparing for cases each class, want PGP office to make groups

Attributes	Levels	Cluster		
		1	2	3
Quant	Low	0.65	-1.01	0.14
	High	-0.65	1.01	-0.14
Pedagogy	Theory based	-0.24	0.51	-0.71
	Project based	0.24	-0.51	0.71
Workload	Even	-0.09	0.54	1.08
	Heavy MT/ET	0.09	-0.54	-1.08
CP	Low	-0.65	-0.25	0.37
	High	0.65	0.25	-0.37
Group Formation	Students	0.57	-0.18	-0.34
	PGP Office/ Faculty	-0.57	0.18	0.34

To summarise the clusters,

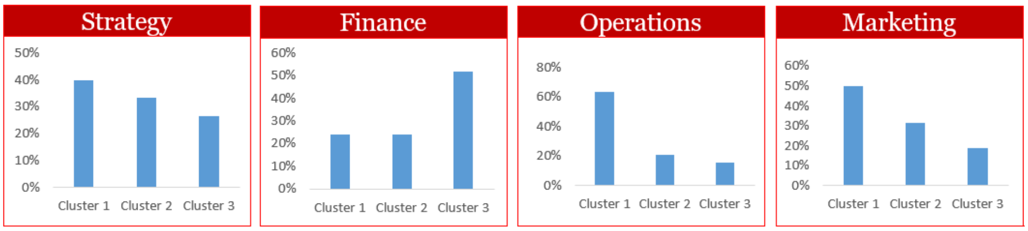
1) The first cluster, comprising of 41% of survey takers, is what we are calling *'The CP Enthusiasts'*. This is because the cluster is characterised by a high preference towards CP (an almost unhealthy level some would say) and an aversion to quant in courses. These people do not really care about workload or pedagogy even, but want to form groups for case/projects themselves. Essentially this can be seen as a group that has confidence in its own CP abilities and wants to be in control of their own fate in courses, assuming that they can pull off some great CP every now and then. The fact that they also want to make their own groups, reinforces this idea of them being in 'control' in a manner of speaking.
2) The second group, with 29% of respondents, are the '*Quant Enthusiasts'*. They have a strong inclination towards courses with high levels of quant content and enjoy theory based teaching. Surprisingly enough, they prefer an even workload as opposed to heavy weightage for mid-terms and end-terms. This could be a pointer for quant heavy courses (like AMDA) for instance. People are much better off with multiple quizzes as evaluative components instead of heavy MT/ETs, to stay on track with the pace of the course.
3) The final cluster is a fairly esoteric cluster, one we are calling '*The Muggus'*, this is because they live to study on a regular basis (demonstrated by their preference for even workload, case based courses) and want the PGP office to make groups for them (perhaps not many people want to work with these people!) The most interesting fact here is the sheer size of the cluster, 30% of survey takers fall into this category (extrapolate it to around 30% of the batch and it would seem that the numbers might very well be correct, 30% of our batch for one would well fit into this category, including some of the authors of this report!)

6. **Inferences:** Of the two tasks that we set out to do as the Academic Council, the first, estimating demand for courses based on their outlines is almost done. We identified three clusters of students who prefer three distinct course outlines. Based on the cluster distribution and extrapolating it to the whole population of students (bear also in mind that the sample was homogenous as we discussed in the beginning and does seem to be a random unbiased sample of the population), here's what we can say,

- A course outline with high CP, low quant, and that allows students to form groups is the most desired. ~40% of the student population will be willing to bid for such courses.
- A course outline with high quant, theory based also will have significant takers. Surprisingly enough, if the course has even workload instead of high MT/ET demand is expected to be higher (~30%).

– For a course which has low CP, but is case based and has even workload with PGP office forming groups, 30% of students can be expected to bid for such a course.

Our second objective of recommending stream-wise optimal course outlines is still undone. Towards this goal, we looked at the 3-cluster distribution for students of each stream. This is given in figure below. This facilitates drawing the following stream-wise inferences,



- **Strategy:** Generally, students prefer Outline 1 (high CP, case based pedagogy, students make groups). However, there is significant demand for the other course outlines too. This seems to signify that the strategy department can do no wrong, in other words, there will always be demand for strategy courses no matter what the outline is like. This is not a luxury that other departments have as the following points will demonstrate.
- **Finance:** Course outlines with even workload and low CP appear to be preferred. Most Finance courses here at IIMB have a high mid-term and/ or high end-term and it would appear that students prefer a smoother evaluative component curve.
- **Operations:** Surprisingly enough, not many people who chose operations prefer high levels of quant in their courses. There could be two reasons here, one, the user base who chose operations seem to have above average work-ex and it could very well be that such people do not prefer quant; two, operations jobs in general might have shifted from a high quant focus to more of a strategy/operations mix focus (also demonstrated by the fact that a lot of these students opted for BOTH operations and strategy). This is also demonstrated in 2nd year Ops. Courses in our opinion, like Supply Chain Management have strong strategy undertones and are fairly limited on quant.
- **Marketing:** While the highest demand is for outlines with high CP, case based courses as is expected, there is also a significant preference for quant-based, low CP courses. This seems to indicate the rising trends of using analytics to marketing problems. Courses like RMD, Digital Marketing etc. are great examples of this.
- **Data Sciences:** Around 10% of our sample opted for Data Sciences courses at all. This is a woeful number and the data science department should look into making their courses more accessible should they wish to increase demand. It could also be a conscious decision to not alter curriculum even if low demand persists and this is fine as well.

7. **Summary:** We started by looking at survey design for the problem in question, which was in essence estimating demand for courses given their outlines. We looked at why survey design is painful, and discussed our travails with good UI design for our survey. We then performed conjoint analysis to get a sense of utilities, and then clustered the respondents into three unique cluster, each cluster preferring a certain kind of course outline. We surmised that this was the solution to our first objective, stating that similar proportion of demand is a good estimate for courses with outlines similar to the ones preferred by each of the three clusters. For our second objective, we divided the data and looked at it stream-wise, drawing inferences from cluster/course outline distribution for each stream.