USE OF SECONDARY DATA IN MARKETING ANALYTICS

A MARKETING ANALYTICS PROJECT

Prepared By,

Debasis Mohanty (UEBA19001)

XIMB – EMBA BUSINESS ANALYTICS

Table of Contents

Introduction:

Marketing analytics is the practice of measuring, managing and analyzing marketing performance to maximize its effectiveness and optimize return on investment. It is a critical part of marketing decision making. It helps in improving the performance by identifying the variable importance measures. Every decision requires relevant information, and strategies which can be developed based on data gathered through marketing analytics. Too often, marketing analytics is considered narrowly as the gathering and analyzing of data for someone else to use. However, firms can achieve and sustain a competitive advantage through the creative use of market information generated by marketing analytics. Marketing analytics is an integral part of marketing research and a bridge to the decision making. Hence, marketing analytics is defined as information input to decisions, not simply the evaluation of decisions that have been made.

Modern marketing analytics is mostly web based on done on internet. It was estimated that by 2023, there would be over 650 million internet users the country. Despite the large base of internet users, the internet penetration rate in the country stood at around 50 percent in 2020. This meant that around half of the 1.37 billion Indians had access to internet that year. As per Marketing Research book by Naresh K. Malhotra, it discussed that the Internet as a source of marketing research information. Analysis of secondary data helps define the marketing research problem and develop an approach. Also, before the research design for collecting primary data is formulated, the researcher should analyze the relevant secondary data. In some projects, particularly those with limited budgets, research may be largely confined to the analysis of secondary data, since some routine problems may be addressed based only on secondary data.

This project report would be about the advantages and disadvantages of secondary data, along with its classification and usages.

Overview of Secondary Data:

Secondary data are the data collected by those who are not related to the research study but for those who have some other purpose in their mind. If the researcher uses these data, then these become secondary data for the current users. A variety of secondary information sources is available to the researcher gathering data on an industry, potential product applications and the marketplace. Secondary data is also used to gain initial insight into the research problem. Secondary data is classified in terms of its source – either internal or external. Internal, or in-house data, is secondary information acquired within the organization where research is being carried out. External secondary data is obtained from outside sources.

Purpose of Secondary Data:

- Model Building: Specifying relationship between two or more variables
- Fact Findings: Descriptive information to support research
- Extracting the relevant information from other sources, previous studies

- Data Mining: Exploring data through computer. Using computer technology to go through volumes of data to discover trends about an organization's sales customers and products. It is primarily used
- Identifying the relevant sources: To avoid plagiarism

Primary Versus Secondary Data:

Primary data are originated by a researcher for the specific purpose of addressing the problem at hand. Obtaining primary data can be expensive and time consuming. The collection of primary data involves all six steps of the marketing research process. They are as follows:

- Defining the Problem
- Developing an Approach to the Problem
- Formulating a Research Design
- Doing Field Work or Data Collection
- Preparing and Analyzing the data
- Preparing and Presenting the Report

Secondary Data are data that have already been collected for purposes other than the problem at hand. These data can be located quickly and inexpensively. The differences between primary and secondary data are summarized in the below table.

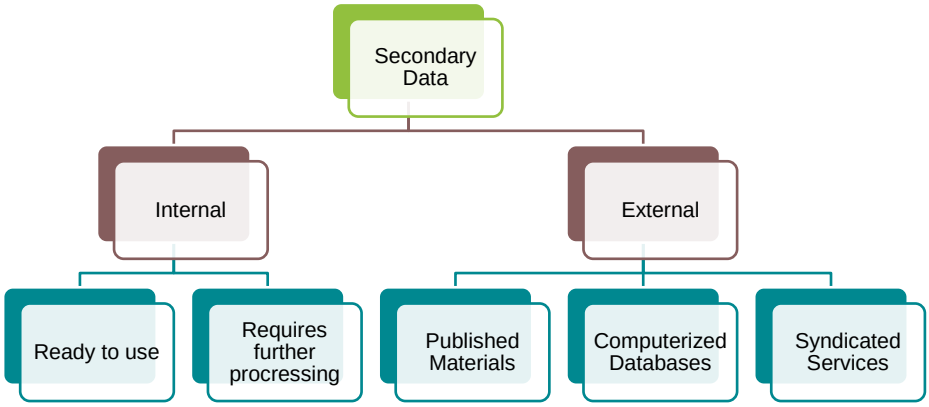
A comparison of Primary and Secondary Data

	Primary Data	**Secondary Data**
Collection purpose	For the problem at hand	For other problems
Collection process	Very involved	Rapid and easy
Collection cost	High	Relatively low
Collection time	Long	Short

Classification of Secondary Data:

Secondary data may be classified as either internal or external. Internal data are those generated within the organization for which the research is being conducted. This information may be available in a ready-to use-format, such as information routinely supplied by the management decision support system. On the other hand, these data may exist within the organization but may require considerable processing before they are useful to the researcher. For example, a variety of information can be found on sales invoices. Yet this information may not be easily accessible; further processing may be required to extract it.

External data are those generated by sources outside the organization. These data may exist in the form of published material, computerized databases, or information made available by syndicated services. Before collecting external secondary data, it is useful to analyze internal secondary data.



Advantages and Uses of Secondary Data:

Secondary data offer several advantages over primary data. Secondary data are easily accessible, relatively inexpensive, and quickly obtained. Although it is rare for secondary data to provide all the answers to a nonroutine research problem, such data can be useful in a variety of ways which is as follows.

- Identify the problem
- Better define the problem
- Develop an approach to the problem
- Formulate an appropriate research design (for example, by identifying the key variables)
- Answer certain research questions and test some hypotheses
- Interpret primary data more insightfully

Given these advantages and uses of secondary data, we state the following general rule:

Examination of available secondary data is a prerequisite to the collection of primary data. Start with secondary data. Proceed to primary data only when the secondary data sources have been exhausted or yield marginal returns.

Secondary data gives a frame of mind to the researcher that in which direction he/she should go for the specific research. Marketing researchers should take advantage of the high-quality secondary data that are available and consider the potential value in gaining knowledge and giving insight into a broad range of marketing issues through utilizing secondary data analysis method.

Limitations and Disadvantages of Secondary Data:

As secondary data has been collected for purposes other than the problem at hand, its usefulness to the current problem may be limited in several important ways, including relevance and accuracy. The following can be the disadvantage of the secondary data:

- Data collected in one location may not be suitable for the other one due variable environmental factor

- With the passage of time the data becomes obsolete and very old
- Secondary data collected can distort the results of the research. For using secondary data, a special care is required to amend or modify for use
- Secondary data can also raise issues of authenticity and copyright

The objectives, nature, and methods used to collect the secondary data may not be appropriate to the present situation. Also, secondary data may be lacking in accuracy, or they may not be completely current or dependable. Before using secondary data, it is important to evaluate them on these factors. These factors are discussed in more detail in the following table.

The major limitation of secondary data analysis of the published data is that, due to the constraints of space due to variable or parameter importance. There is considerable information in the original data that cannot be recovered from the summary measures reported in the published article. If the original data were available, this limitation of secondary data analysis would disappear.

Criteria for Evaluating Secondary Data:

Criteria for Evaluating Secondary Data

Criteria	Issues	Remarks
Specifications/ Methodology	Data collection method Response rate Quality of data Sampling technique Sample size Questionnaire design Fieldwork Data analysis	Data should be reliable, valid, and generalizable to the problem at hand
Error/Accuracy	Examine errors in approach, research design, sampling, data collection, data analysis, reporting	Assess accuracy by comparing data from different sources
Currency	Time lag between collection and publication Frequency of updates	Census data are periodically updated by syndicated firms
Objective	Why were the data collected?	The objective will determine the relevance of the data
Nature	Definition of key variables Units of measurement Categories used Relationships examined	Reconfigure the data to increase their usefulness, if possible
Dependability	Expertise, credibility, reputation, and trustworthiness of the source	Data should be obtained from an original rather than an acquired source

The quality of secondary data should be routinely evaluated, using the criteria present in the above table. The specifications or the methodology used to collect the data should be critically examined to identify possible sources of bias. Such methodological considerations include size and nature of the sample, response rate and quality, questionnaire design and administration, procedures used for fieldwork, and data analysis and reporting procedures. These checks provide information on the reliability and validity of the data and help determine whether they can be generalized to the problem at hand. The reliability and validity can be further ascertained by an examination of the error, currency, objectives, nature, and dependability associated with the secondary data.

Usage of Secondary Data Sources in Marketing:

With secondary research in hand, the next step is to review your source materials to pull out the insights that are most pertinent to your marketing problem. Some secondary research sources may include data you can analyze and map to your own customer segmentation or other market analyses. Other secondary research provides analysis and insights you can use to develop implications and recommendations for your organization and marketing problem.

It is helpful to capture key findings and recommendations from the secondary research review and analysis, just as you would for a primary research project. The goal is to summarize what you have learned, making it easier for any primary research activity to build on what has already been discovered from secondary research.

Secondary Data Sources that are Public:

There are several data sources that are public and most of the data is uploaded here are secondary data sets for a specific purpose and analysis. A few top great sites with free data sets are listed below:

- Kaggle (https://www.kaggle.com/)
- GitHub (https://github.com/)
- UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)
- Openml (https://www.openml.org/)
- Reddit (https://www.reddit.com/r/datasets/)
- Quandl (https://www.quandl.com/)
- U.S. Government's open data (https://www.data.gov/)
- India Government's open data (https://data.gov.in/)

There are many more online websites where the dataset repository is created by few book publishing houses, universities and author's private blogs to support consumers in using the datasets which is used in the book by the author. Moreover, few authors use GitHub as there repository as well. For example, for the book named "*Basic Practice of Statistics, 7th edition*" the data files are kept at https://www.austincc.edu/mparker/software/data/

Secondary Data Sources that are Purchased:

These are those data which are used in industry or have a non-disclosure agreement between two parties. Moreover, data which is copyright protected, company classified data, stock market data like Bloomberg and online competition or hackathon data are some of the secondary datasets which must be purchased as it has high value insights. To name a few sites:

- Bloomberg (https://www.bloomberg.com/professional/solution/bloomberg-terminal/)
- Harvard Business School Datasets used in Education Institutions (https://hbsp.harvard.edu/search?N=4294930434+516162&Ns=publication_date_filter%7C1%7C%7Caggregate_sort%7C0&action=sort)
- Stock Market datasets used by traders and researchers (https://www.nseindia.com/market-data/analytical-products)

How Secondary Data is used:

For a primary analysis, the investigator must select summary measures to report in the text, figures, tables, and short appendices. The decision about how to summarize the data is an important one, because it is irreversible. It is seldom possible to regenerate the original data from the summaries and, therefore, the published summary measures of performance can only rarely be used to examine alternative measures of performance. For example, a study of classical conditioning that reports absolute or relative rates (or probabilities) of responding in the presence and absence of a stimulus cannot be used to evaluate a real-time theory of conditioning because the time of responding has been eliminated from the record.

Meta-Analysis is performed. Meta-analysis has become an important research strategy as it enables researchers to combine the results of many pieces of research on a topic to determine whether the findings holds generally. This is better than trying to assume that the findings of a single study have global meaning.

Meta-Analysis is an objective and quantitative methodology for synthesizing previous studies and research on a particular topic into an overall finding.

The term meta-analysis means '*an analysis of analysis*'. A particular topic may have been replicated in various ways, using, for example, differently sized samples, and conducted in different countries under different environmental, social and economic conditions. Sometimes results appear to be reasonably consistent; others less so. Meta-analysis enables a rigorous comparison to be made rather than a subjective 'eyeballing'. However, the technique relies on all relevant information being available for each of the examined studies. If some crucial factors like sample size and methodology are missing, then comparison is not feasible.

Meta-analysis allows us to compare or combine results across a set of similar studies. In the individual study, the units of analysis are the individual observations. In meta-analysis the units of analysis are the results of individual studies.

There are three stages to this:

- Identify the relevant variables:
  This sounds easy, but like defining a hypothesis and determining research questions, you must be specific and clear about what your real focus is. You cannot simply say, 'I want to do a meta-analysis on attitude change research'. So, you must limit yourself to conducting an evaluation of a much smaller segment, such as the effects of different types of feedback on job performance, or the effects of different levels of autonomy on group achievement of objectives in the service industry.

- Locate relevant research:
  One issue that is vital and potentially serious for the meta-analyst is the file drawer problem. The file drawer phenomenon is potentially serious for meta-analysis because it produces a biased sample – a sample of only those results published because they reported statistically significant results. This bias inflates the probability of making a Type II error (concluding that a variable has an effect when it does not). Studies that failed to be published are not available to be included in the meta-analysis. Because meta-analytic techniques ultimately lead to a decision based on available statistical information, an allowance must be made for the file drawer phenomenon. There are two ways of dealing with the file drawer problem.

  First, uncover those studies that never reach print by identifying as many researchers as possible in the research area you are researching. Then send each a questionnaire, asking if any unpublished research on the issue of interest exists. This may be impracticable in some topics as identification of researchers is difficult and non-response may be high anyway. But most of the secondary data is get generated out of structured datasets.

  A second but more practical approach, involves calculating the number of studies averaging null results (i.e. that did not reach significance) that would be required to push the significance level for all studies, retrieved and unretrieved combined, to the 'wrong' side of significance level ($p = 0.05$).

- Conduct the meta-analysis:
  When you have located relevant literature, collected your data, and are reasonably certain that the file drawer phenomenon isn't an issue, you are ready to apply one of the many available meta-analytic statistical techniques. The heart of meta-analysis is the statistical combination of results across studies. Therefore, as well as recording the methodology, sample, design, hypotheses, conclusions, etc., you must record information particularly from the results section of research papers you are reviewing such a r's, t's, chi squares, F's, and p values.

Apart from Meta-Analysis there are other analysis of secondary data which is performed by many people to extract deep insights for decision making. Some of them are:

- Cluster Analysis
- Predictive Analysis
- Discriminant Analysis
- Anomalies Detection
- Business Analysis for Decision Making

Who all uses Secondary Data:

Secondary data is used by many people across the domain. Once the raw data is generated in an initial analysis or meta-analysis phase many people are engaged across the domain in data preprocessing and data cleansing activities. Most of the data is normalized and standardized (feature scaling) to make sense which is then further tested using statistical tools and hypothesis. In fact, many researchers, students, data analyst, data scientist rely on secondary data to prepare exit report and use it in publishing their research paper. The higher management and stakeholders mostly deal with the secondary data as these datasets are the ones which can provide them relevant insights.

Real-Life Application:

Below is example of GRE Dataset analysis for graduate admission. I have taken this dataset from Kaggle where admission prediction is calculated based on GRE Score, TOEFL Score, University Rating, Statement of purpose ratings, Letter of recommendation ratings, CGPA, Research criteria and Percentage chance of Admission. Now, why I have taken this dataset because I am curious to know who all joins the premier colleges which what scores and its variable importance.

Example:

Below are the top 4 records. Please check the reference section (the last point) for the link to download the dataset on your machine.

Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
1	337	118	4	4.5	4.5	9.65	1	0.92
2	324	107	4	4	4.5	8.87	1	0.76
3	316	104	3	3	3.5	8	1	0.72
4	322	110	3	3.5	2.5	8.67	1	0.8

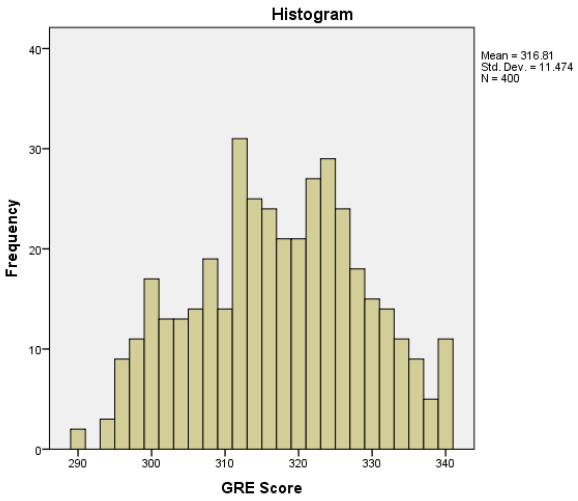
Exploratory Data Analysis:

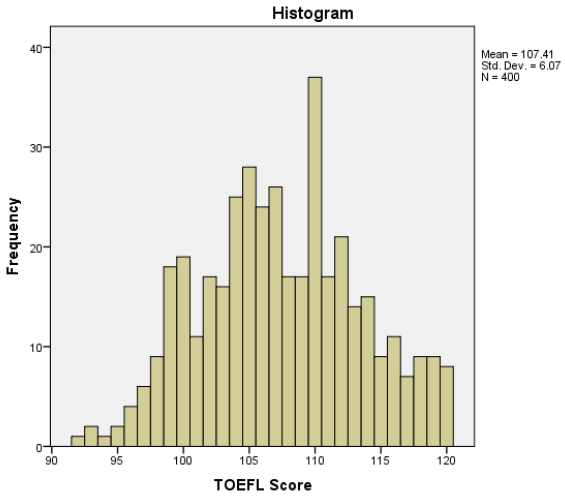
Before we run any analysis (like discriminant or cluster), we need to review the assumptions by exploring the dataset.

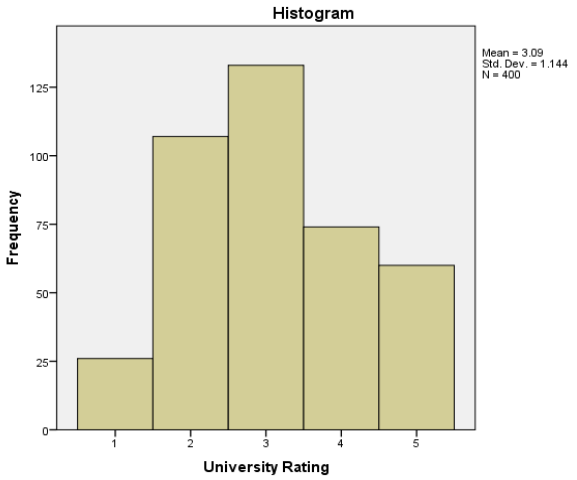
- The predictors must be independent
- Group member must be mutually exclusive. For example, a student can't be admitted and not admitted at the same time
- There must be an absence or negligible outliers
- The predictors should be normally distributed. And the in-group variance-variance matrix should be equal across groups.
- Also, there should not be any multicollinearity issue between the variables in the dataset
- The categorical variables in the dataset should be balanced else its effect might get nullified.

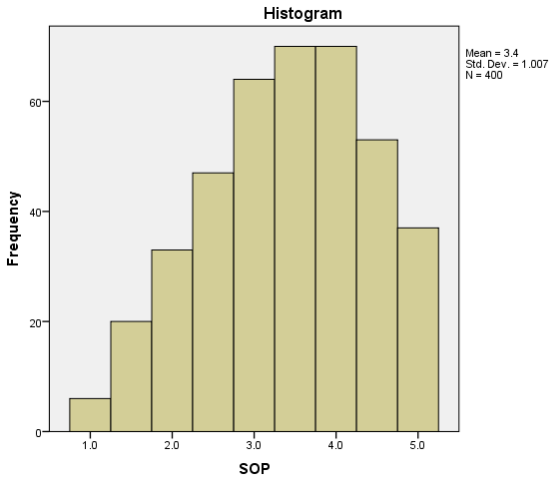
Case Processing Summary

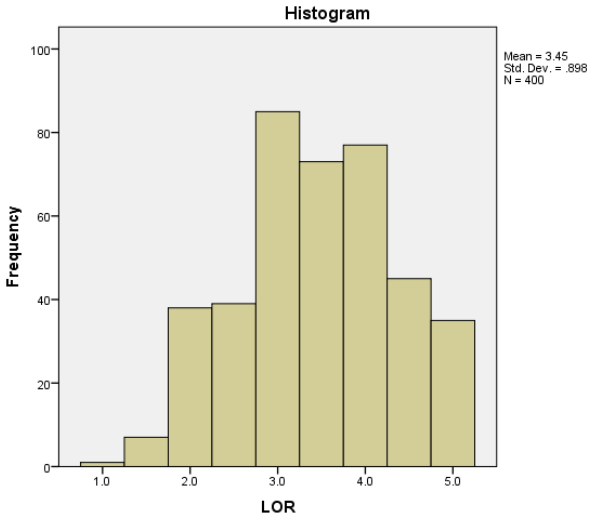
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
GRE Score	400	100.0%	0	0.0%	400	100.0%
TOEFL Score	400	100.0%	0	0.0%	400	100.0%
University Rating	400	100.0%	0	0.0%	400	100.0%
SOP	400	100.0%	0	0.0%	400	100.0%
LOR	400	100.0%	0	0.0%	400	100.0%
CGPA	400	100.0%	0	0.0%	400	100.0%
Research	400	100.0%	0	0.0%	400	100.0%
Chance of Admit	400	100.0%	0	0.0%	400	100.0%

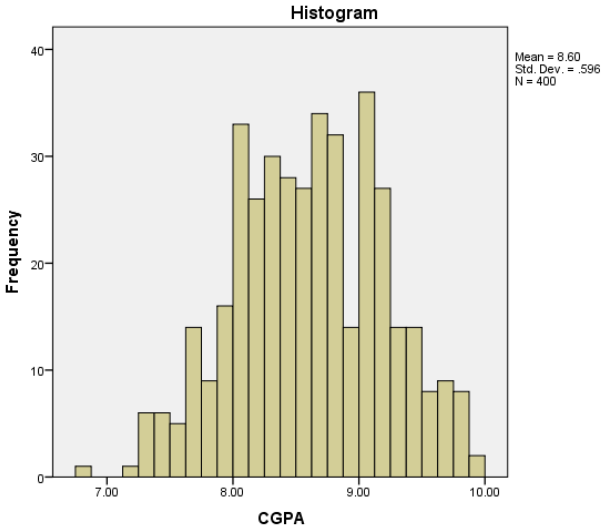


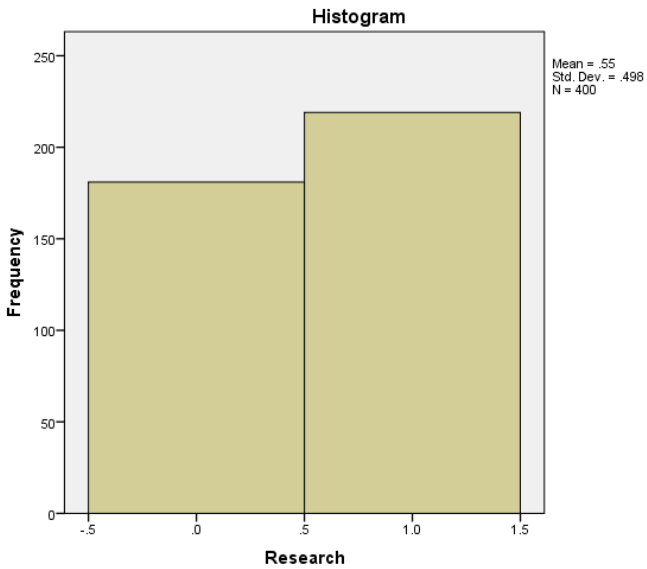


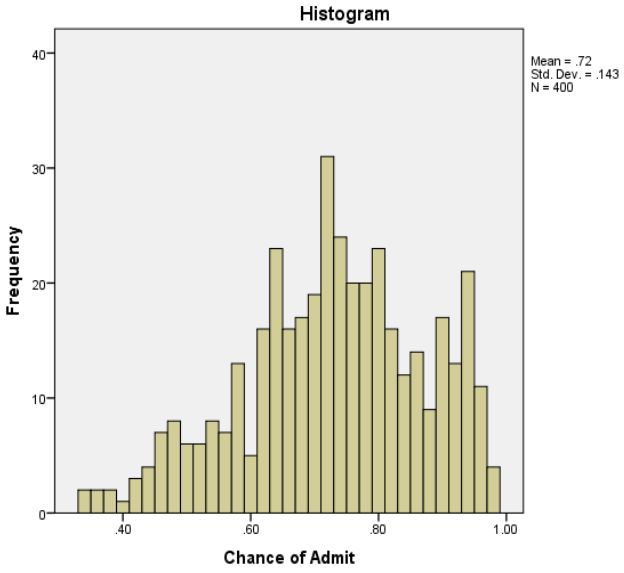

Histogram
Mean = 3.09
Std. Dev. = 1.144
N = 400
Frequency
University Rating


Histogram
Mean = 3.4
Std. Dev. = 1.007
N = 400
Frequency
SOP


Histogram
Mean = 3.45
Std. Dev. = .898
N = 400
Frequency
LOR


Histogram
Mean = 8.60
Std. Dev. = .596
N = 400
Frequency
CGPA


Histogram
Mean = .55
Std. Dev. = .498
N = 400
Frequency
Research


Histogram
Mean = .72
Std. Dev. = .143
N = 400
Frequency
Chance of Admit

We could see that there is no missing data and the Descriptive Statistics for each variable being generated by SPSS (Kindly check for the attached SPSS files for the same analysis results). Moreover, we can understand that many variables are normally distributed. This shows that we can go ahead with data analysis techniques like clustering and logistic regression. We have also identified that we have outliers present in variables like LOR, CGPA and Chance of Admit.

Descriptives

			Statistic	Std. Error
GRE Score	Mean		316.81	.574
	95% Confidence Interval for Mean	Lower Bound	315.68	
		Upper Bound	317.94	
	5% Trimmed Mean		316.80	
	Median		317.00	
	Variance		131.645	
	Std. Deviation		11.474	
	Minimum		290	
	Maximum		340	
	Range		50	
	Interquartile Range		17	
	Skewness		-.063	.122
	Kurtosis		-.700	.243

Please check the Descriptive Statistics for other variables in the SPSS Output file named *Exploratory_Data_Analysis*.

Cluster Analysis:

Further we have observed that there are no such categorical decision variable present in the dataset. Looking at the dataset and above histogram we can predict that the percentage Change of Admit needs to be binned into categories to create decision variables. Before binning let us run the Two Step clustering algorithm to understand the model summary and cluster quality.

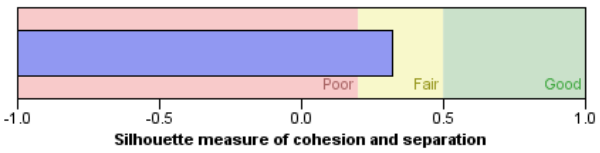
Research

		0		1	
		Frequency	Percent	Frequency	Percent
Cluster	1	79	43.6%	71	32.4%
	2	1	0.6%	114	52.1%
	3	101	55.8%	34	15.5%
	Combined	181	100.0%	219	100.0%

Model Summary

Algorithm	TwoStep
Inputs	8
Clusters	3

Cluster Quality



Now we can confirm that we have 3 clusters based on the cluster quality. So, using the Chance of Admit variable and categorizing into 3 status we have. High Chance, Medium Chance and

Low Chance of getting admission. High Chance being > 80%, Low Chance < 60% and the rest being Medium Chance.

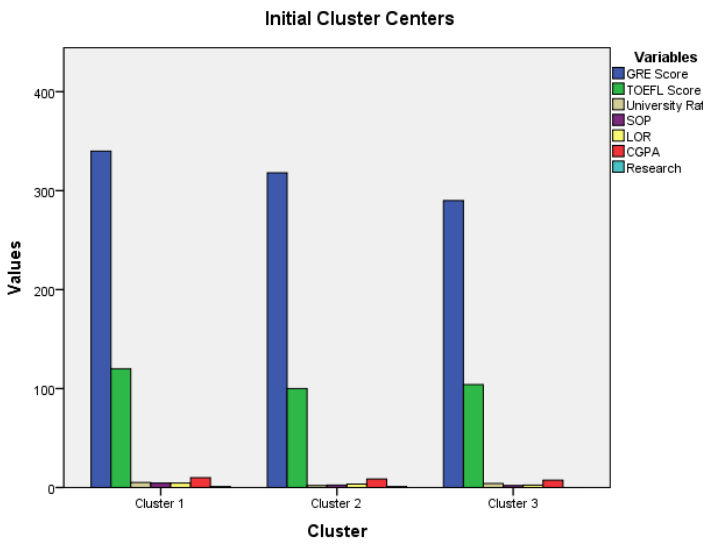
Performing K Means Clustering we have the following output with all variable being significant in the ANOVA table.

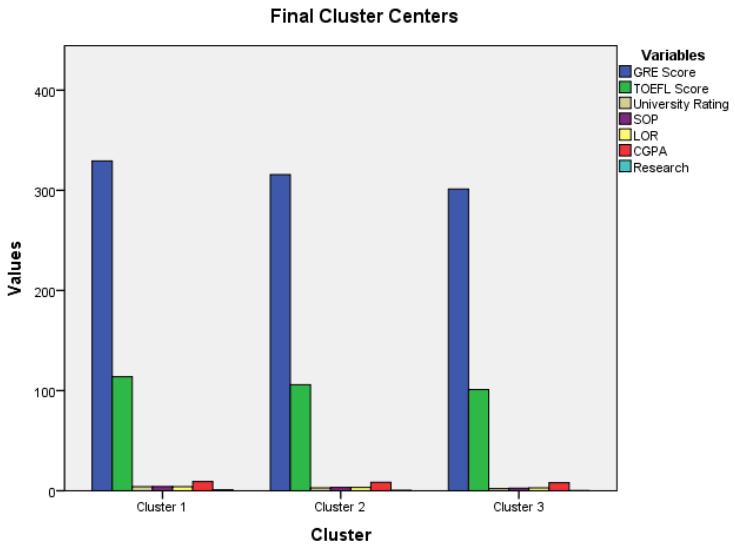
Initial Cluster Centers

	Cluster		
	1	2	3
GRE Score	340	318	290
TOEFL Score	120	100	104
University Rating	5	2	4
SOP	4.5	2.5	2.0
LOR	4.5	3.5	2.5
CGPA	9.91	8.54	7.46
Research	1	1	0

Final Cluster Centers

	Cluster		
	1	2	3
GRE Score	329	316	301
TOEFL Score	114	106	101
University Rating	4	3	2
SOP	4.1	3.3	2.6
LOR	4.0	3.4	2.8
CGPA	9.19	8.48	7.98
Research	1	0	0



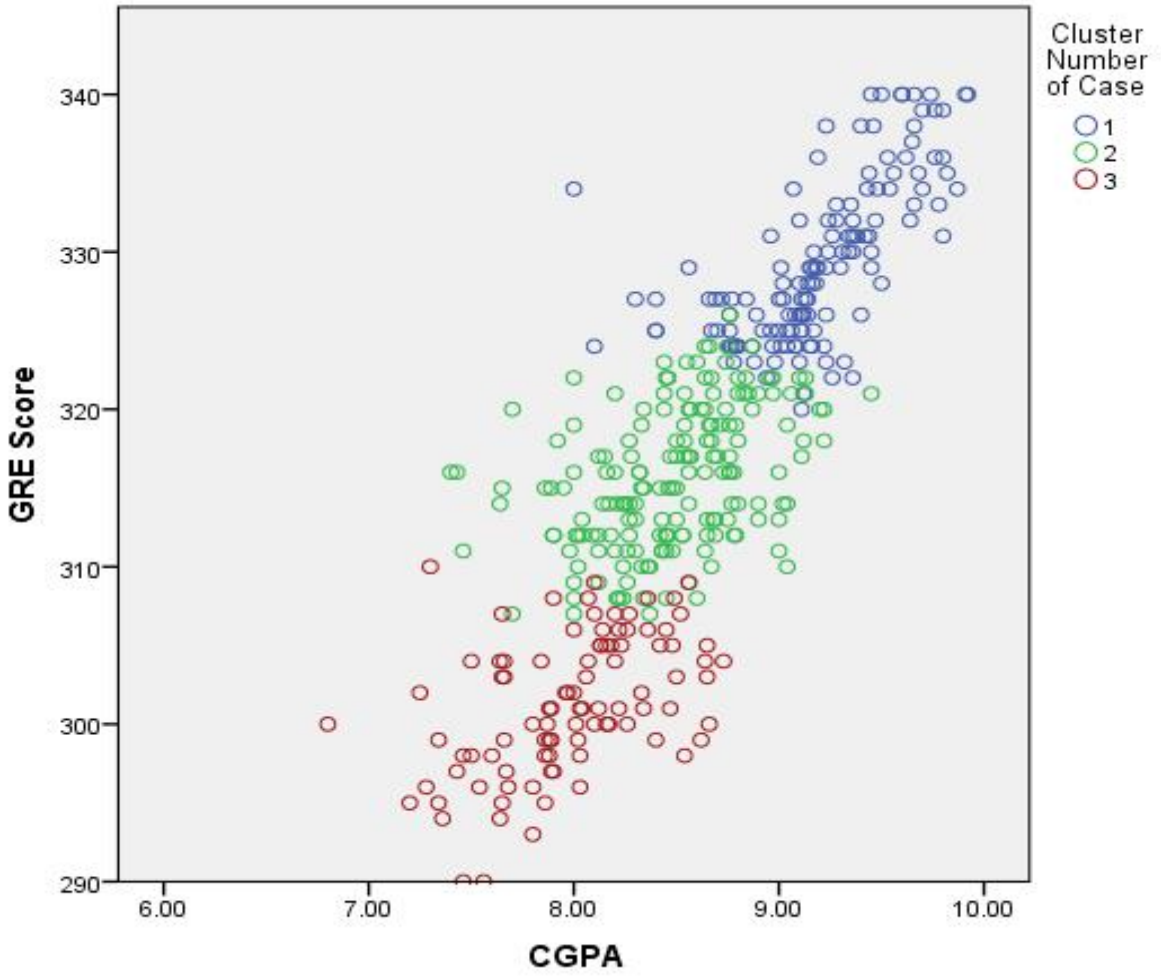


Cluster Analysis is the statistical method of partitioning a sample into homogeneous classes to produce an operational classification.

Because we usually don't know the number of groups or clusters that will emerge in our sample and because we want an optimum solution, a two-stage sequence of analysis occurs as follows:

- We carry out a Two Step cluster analysis using Ward's method applying squared Euclidean Distance as the distance or similarity measure. This helps to determine the optimum number of clusters we should work with.
- The next stage is to run other clustering techniques like the K Mean cluster analysis with our selected number of clusters, which enables us to allocate every case in our sample to a particular cluster.

This sequence and methodology using SPSS will be described in more detail later. There are a variety of clustering procedures of which hierarchical cluster analysis is the major one.

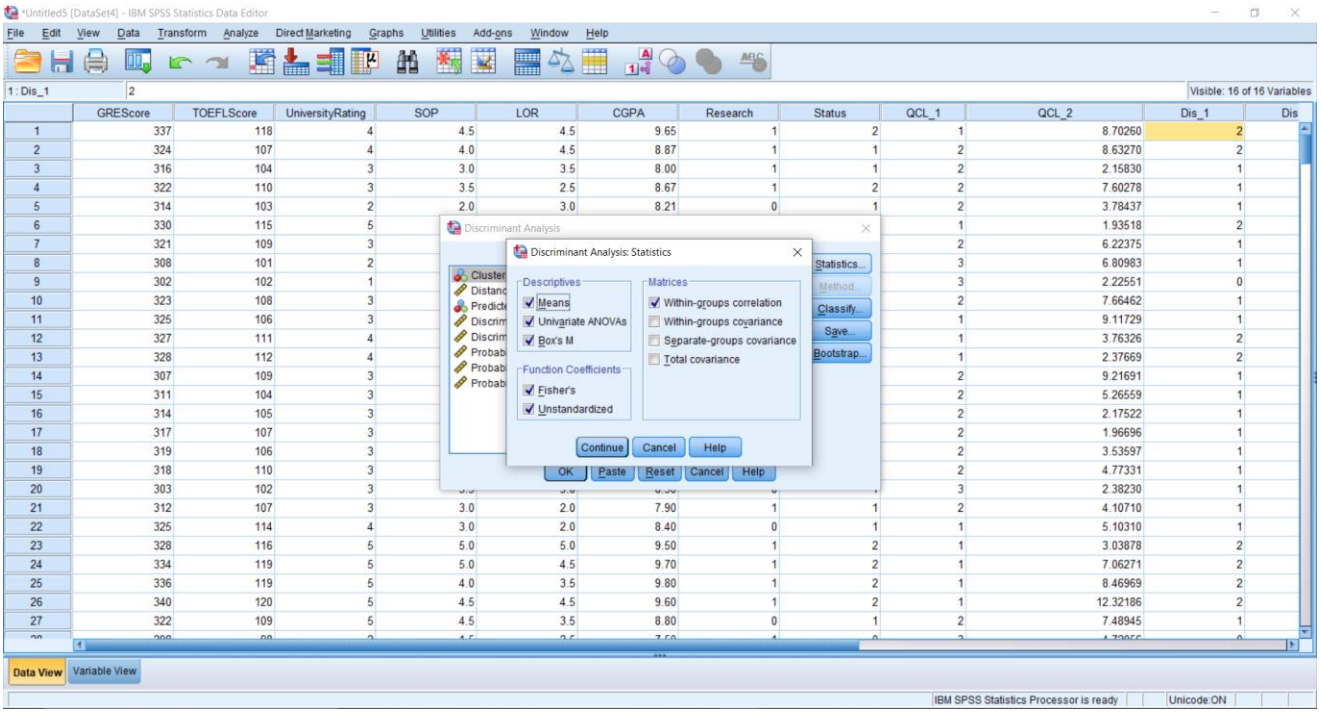


Here Cluster 1 being the High Chance, Cluster 2 as Medium Chance and the Cluster 3 as Low Chance. I have plotted few other cluster graphs in the SPSS output for better understanding of the data.

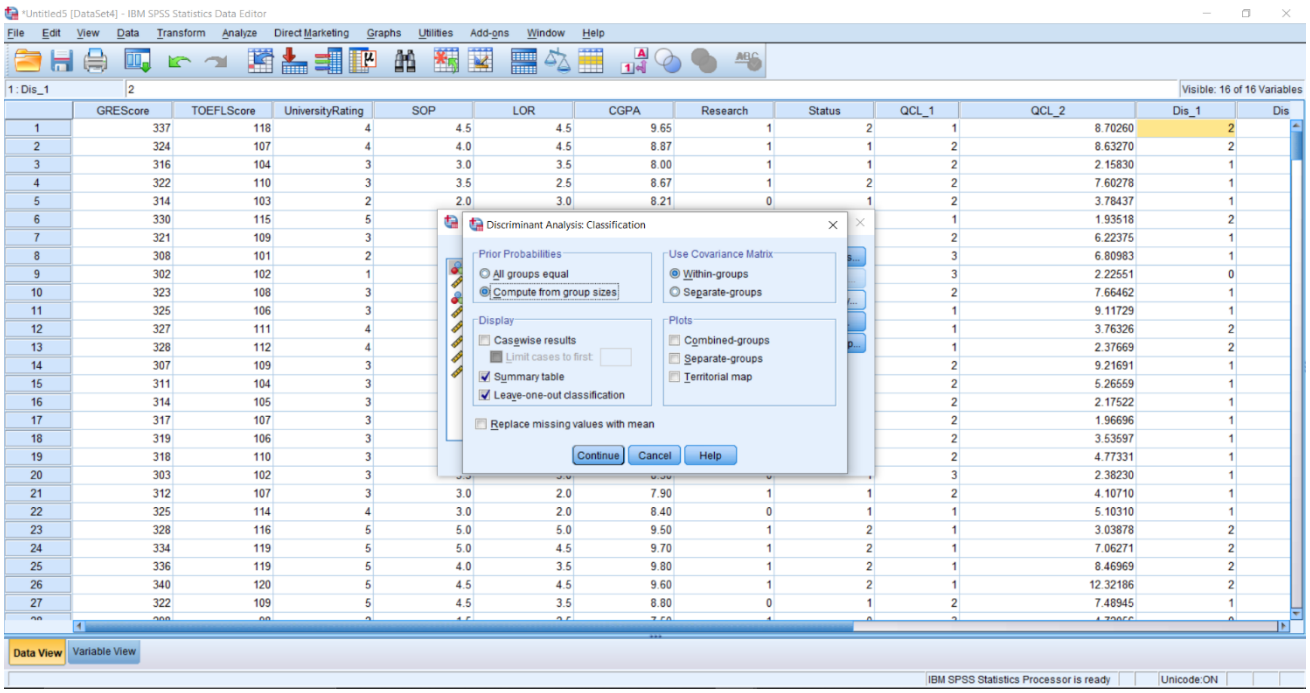
Discriminant Analysis:

Discriminant Function Analysis (DA) undertakes the same task as multiple linear regression by predicting an outcome. However, multiple linear regression is limited to cases where the dependent variable on the Y axis is an interval variable so that the combination of predictors will, through the regression equation, produce estimated mean population numerical Y values

for given values of weighted combinations of X values. But many interesting variables are categorical, such as Chance of getting admitted in this case analysis.



In SPSS, we will calculate the within-group statistics with function coefficient being checked for both Fisher's and Unstandardized. Moreover, as the number of group sizes are not same across the dataset, we will use prior probabilities to be calculated based on group size as per below SPSS screenshot.



Discriminant Analysis SPSS Report

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
GRE Score	.438	255.147	2	397	.000
TOEFL Score	.433	259.859	2	397	.000
University Rating	.527	178.399	2	397	.000
SOP	.572	148.728	2	397	.000
LOR	.626	118.377	2	397	.000
CGPA	.329	405.413	2	397	.000
Research	.695	87.261	2	397	.000

Pooled Within-Groups Matrices

		GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research
Correlation	GRE Score	1.000	.624	.319	.248	.199	.578	.305
	TOEFL Score	.624	1.000	.372	.339	.219	.564	.137
	University Rating	.319	.372	1.000	.523	.427	.441	.114
	SOP	.248	.339	.523	1.000	.552	.422	.141
	LOR	.199	.219	.427	.552	1.000	.380	.105
	CGPA	.578	.564	.441	.422	.380	1.000	.153
	Research	.305	.137	.114	.141	.105	.153	1.000

Box’s Test of Equality of Covariance Matrices

Test Results

Box's M		184.784
F	Approx.	3.201
	df1	56
	df2	175151.554
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

Log Determinants

Status	Rank	Log Determinant
Low Chance	7	-.256
Medium Chance	7	-.256
High Chance	7	-3.519
Pooled within-groups	7	-.834

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Box’s M is 184.784 with F = 3.201 which is significant at p < .000

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2.496[a]	98.7	98.7	.845
2	.032[a]	1.3	100.0	.175

a. First 2 canonical discriminant functions were used in the analysis.

A canonical correlation of .845 suggests the model explains 71.40% of the variation in the grouping variable, i.e. whether a candidate chance of admission is more or not.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.277	505.458	14	.000
2	.969	12.277	6	.056

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
GRE Score	.109	-.155
TOEFL Score	.214	.497
University Rating	.145	.601
SOP	.051	-.447
LOR	.085	-.611
CGPA	.567	-.306
Research	.235	.496

Canonical Discriminant Function Coefficients

	Function	
	1	2
GRE Score	.014	-.020
TOEFL Score	.053	.124
University Rating	.175	.723
SOP	.067	-.585
LOR	.119	-.857
CGPA	1.653	-.891
Research	.565	1.191
(Constant)	-25.987	2.849

Unstandardized coefficients

Functions at Group Centroids

Status	Function	
	1	2
Low Chance	-2.253	.272
Medium Chance	-.514	-.169
High Chance	2.097	.105

Unstandardized canonical discriminant functions evaluated at group means

Structure Matrix

	Function	
	1	2
CGPA	.904*	-.194
TOEFL Score	.724*	.234
GRE Score	.717*	.089
University Rating	.600*	.164
SOP	.546*	-.398
Research	.417*	.411
LOR	.484	-.586*

Classification Statistics

Prior Probabilities for Groups

Status	Prior	Cases Used in Analysis	
		Unweighted	Weighted
Low Chance	.185	74	74.000
Medium Chance	.495	198	198.000
High Chance	.320	128	128.000
Total	1.000	400	400.000

Classification Function Coefficients

	Status		
	Low Chance	Medium Chance	High Chance
GRE Score	5.833	5.867	5.899
TOEFL Score	.519	.557	.731
University Rating	-14.127	-14.143	-13.488
SOP	-3.695	-3.320	-3.306
LOR	.776	1.362	1.439
CGPA	12.363	15.632	19.704
Research	-29.497	-29.039	-27.237
(Constant)	-944.747	-987.801	-1057.368

Fisher's linear discriminant functions

Classification Results[a,c]

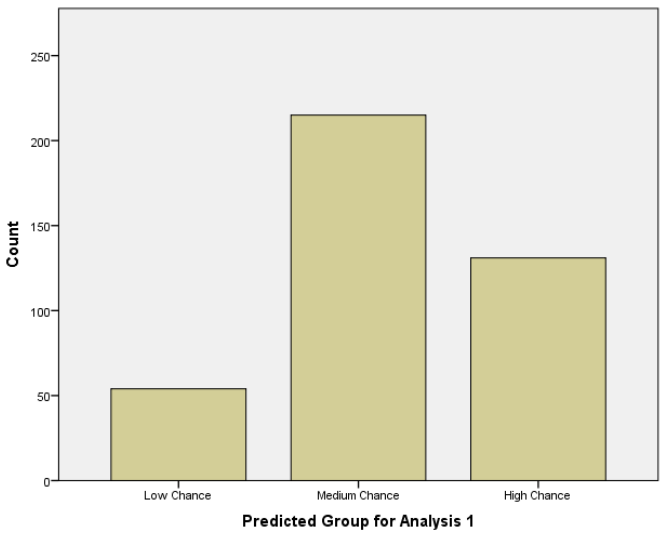
			Predicted Group Membership			
		Status	Low Chance	Medium Chance	High Chance	Total
Original	Count	Low Chance	46	28	0	74
		Medium Chance	8	175	15	198
		High Chance	0	12	116	128
	%	Low Chance	62.2	37.8	.0	100.0
		Medium Chance	4.0	88.4	7.6	100.0
		High Chance	.0	9.4	90.6	100.0
Cross-validated[b]	Count	Low Chance	44	30	0	74
		Medium Chance	12	169	17	198
		High Chance	0	13	115	128
	%	Low Chance	59.5	40.5	.0	100.0
		Medium Chance	6.1	85.4	8.6	100.0
		High Chance	.0	10.2	89.8	100.0

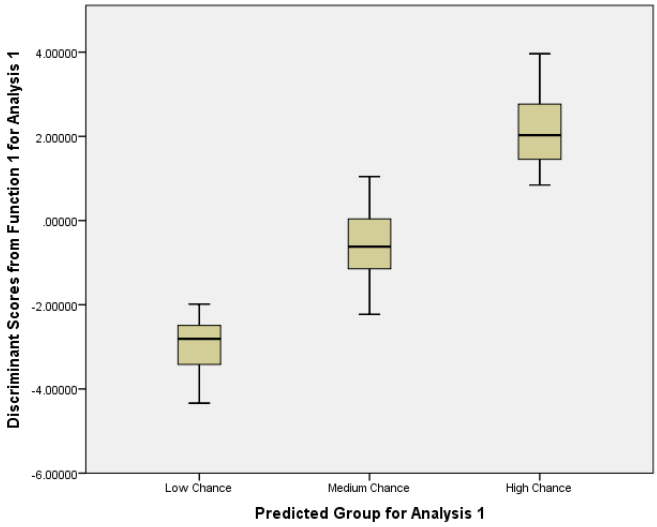
Summary of Discriminant Analysis:

Wilks' lambda indicates the significance of the discriminant function. This table indicates a highly significant function ($p < .000$) and provides the proportion of total variability not explained, i.e. it is the converse of the squared canonical correlation. So, we have 22.7% unexplained.

From the Canonical Discriminant Function Coefficients table these unstandardized coefficients are used to create the discriminant function (equation)

$$D = (0.014 \times GRE\ Score) + (0.053 \times TOEFL\ Score) + (0.175 \times University\ Rating) + (0.67\ SOP) + (.119\ LOR) + (1.653\ CGPA) + (0.565\ Research) - 25.987$$





Finally, there is a classification phase. The above classification table, also called a confusion table, is simply a table in which the rows are the observed categories of the dependent and the columns are the predicted categories. When prediction is perfect all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications. The cross validated set of data is a more honest presentation of the power of the discriminant function than that provided by the original classifications and often produces a poorer outcome. Furthermore, the cross validation produces a more reliable function. 82.0% of cross-validated grouped cases correctly classified. The argument behind it is that one should not use the case you are trying to predict as part of the categorization process.

The classification results reveal that 84.3% of respondents were classified correctly into 'High', 'Medium" or 'Low' chance groups. This overall predictive accuracy of the discriminant function is called the 'hit ratio'. The High chance is predicted at a better accuracy 90.6% than Medium (88.4%) and Low (62.2%).

My Experience:

As a part of corporate restructuring, merger and acquisition I worked for British Telecom as a contractor employee on behalf of my company TCS. As a Data Analyst I was a part of Data cleansing team for asset migration activities. On the floor, I was dealing with a deluge of computer-generated data however, there were lot of data discrepancies and redundancy in the data which entire team used to cleanse it before sending it to the stakeholders. We also used to perform similar analysis as stated above however they were more related to telecom domain. It was a nice experience in dealing with a lot of secondary then but now I realize the real use of it and could connect the dots as I figured out from where those data originated initially. I won't be able to discuss much about the data which I used due to non-disclosure agreement, but I can say that it really makes a lot sense to me. This really clarify the difference between the primary and secondary data.

References:

The below references are being used to make this project:

- The Effective Use of Secondary Data, Journal published in Elsevier. https://www.sciencedirect.com/science/article/abs/pii/S0023969001910987
- Accessing Secondary Data : A Literature Review, a journal from ResearchGate by Kalu, Alexanda, Larry Chukwuemeka Unachukwu, Oti Ibiam (https://www.researchgate.net/publication/328067351_Accessing_Secondary_Data_A_Literature_Review)
- Discriminant Analysis Chapter 25 published by Sage, chapters from a book name "*Business Research Methods and Statistics using SPSS*" by Robert B. Burns and Richard A. Burns. https://github.com/DragonflyStats/AdvancedStatistics/blob/master/Chapter%2025%20-%20Discriminant%20Analysis.pdf
- Marketing Research Book by Naresh K. Malhotra, an applied orientation textbook.
- Class Notes and PPT used in our Marketing Analytics class by Plavini Punyatoya for Discriminant Analysis and Cluster Analysis
- Dataset from Kaggle, GRE Dataset analysis for graduate admission. The Admission_Predict.csv dataset. https://www.kaggle.com/mohansacharya/graduate-admissions