SUMMER TRAINING REPORT

On

MINING AND ANALYSIS OF INSURANCE POLICIES CUSTOMER DATABASE

Submitted to Guru Gobind Singh Indraprastha University, Delhi (India) in partial fulfillment of the requirement for the award of degree of

Bachelor of Technology

In

Information Technology

Submitted by:

PRANAV GUPTA
Roll No: 03196303114

Department of Information Technology

Maharaja Surajmal Institute of Technology,
New Delhi - 110058

(2014-2018)

ACKNOWLEDGEMENT

I, Pranav Gupta, of Maharaja Surajmal Institute of Technology, pursuing B.Tech in IT, am extremely grateful to Landmark Insurance Brokers Pvt. Ltd. for the confidence bestowed in me and entrusting my project entitled "Mining and Analysis of Insurance Policies Customer Database".

At this juncture I feel deeply honored in expressing my sincere thanks to the Project Mentor, Mr. Rajendra Rawal for making the resources available at right time and providing valuable insights leading to the successful completion of the project.

I express my gratitude to our college Director Prof. B.S Panwar and our department H.O.D Mr. Manoj Malik for arranging the summer training in good schedule.

I would also like to thank all the faculty members of Maharaja Surajmal Institute of Technology for their critical advice and guidance without which this project would not have been possible.

Submitted by:
PRANAV GUPTA
03196303114

Candidate's Declaration

I, Pranav Gupta, enrollment number 02996303114, B.Tech in IT (5th Semester) of Maharaja Surajmal Institute of Technology, New Delhi, hereby declare that the Training Report titled "Mining and Analysis of Insurance Policies Customer Database" is an original work and data provided in the study is authentic to the best of my knowledge. This report has not been submitted to any other institute for the award of any degree.

Place: New Delhi	Pranav Gupta
Date: 05/09/2016	(03196303114)

ABSTRACT

STRATEGIC IMPORTANCE OF DATA MINING AND ANALYSIS-

The main challenge for companies is to identify accurate models and methods to predict winning competitive strategies. Data mining is becoming an astonishing approach for data analysis because the meaningful knowledge is often hidden in enormous databases, and most traditional statistical methods could fail to uncover such knowledge. An efficient development of the customer relationship management and the data mining is the vital resource to collect and to manage this knowledge.

This project demonstrates the strong relationship between data mining and customer relationship management in order to forecast customer-centric marketing strategies. It also shows the results of an empirical study related to the identification of the main marketing and financial activities that could be leading customers in a credit-risk state.

With Data Mining, companies can make better and more effective business decisions – marketing, advertising, etc – decisions that will help these companies grow.

USES OF DATA ANALYTICS

Data analytics is the process of examining large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits.

Data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. Mainstream Business Intelligence software and data visualization tools can also play a role in the analysis process.

ORGANISATION INTRODUCTION

Landmark Insurance Brokers Pvt. Ltd. is one of India's leading insurance sources, with PolicyBoss.com being its online presence. At PolicyBoss.com, you can find all the insurances you could need: Car, Health, Term, Travel, Home, Office and Marine. They are also India’s leading motor insurance broker in the retail segment. In the 13 years since its inception in 2003, Landmark has built a strong footprint in the Indian insurance industry.

Today, Landmark caters to more than 3 Lac customers. It is partnered with more than 50 insurance service providers and has a full-fledged presence in 22 cities. This nationwide chain of offices, staffed by over 1000 trained experts, makes buying insurance and registering a claim a convenient pleasure. They are the only broker who helps customers with their claims and they even help if the customer did not buy the claim from them.

In just 3 years of going online, PolicyBoss.com's track record of providing free and competitive online insurance quotes has created a benchmark of its own. The user-friendly site helps you compare policy options easily, quickly and effortlessly buy the best insurance plan in a few simple steps.

MISSION:

To be India's broker of choice for innovative insurance products, with a reputation for quality of service and speed of implementation and execution.

To be among India's leading employers, with an exclusive skill to discover talent, offer plenty of opportunities and ensure that their careers will blossom as their potential is recognized, developed and harnessed.

To be a profitable partner to clients and insurance companies.

To maintain the track record of conducting every relationship with fairness, transparency and a 'win-win' outcome for all involved.

Landmark Insurance ®

SOLUTIONS. NOT JUST ANSWERS.

Date: 27th Jul'16

TO WHOM IT MAY CONCERN

This is to certify that Mr. Pranav Gupta, a student of B. Tech. (IT), Enrollment No: 03196303114, Maharaja Surajmal Institute of Technology, IP University, Delhi, India has successfully completed 6-week (From 13th Jun'16 to 26th Jul'16) long internship program at our Mumbai Branch. During the period of his internship with us he was found punctual, hardworking and inquisitive.

We wish him every success in life.

Yours Faithfully

Didwania

Ms. Disha Didwania

Assistant Manager – HR

Landmark Insurance Brokers Pvt. Ltd.
(CIN: U66030PN2002PTC017009)
[IRDA License No: 216 | Broker Code: DB 089/03| Valid till 12/10/2018| Direct (Life & General]
Regd. Office: Unit No. 402, 4th Floor, Business Avenue, Lane No.6, Koregaon Park, Pune (Maharashtra) – 411001
Corporate Office: Unit No.3, 4th Floor, The Centrium, Phoenix Marketcity, LBS Marg, Kurla (West), Mumbai (Maharashtra) – 400 070
Email :solutions@policyboss.com | Website: www.policyboss.com | Toll Free No:18004194199.
Insurance is subject matter of solicitation

MAHARAJA SURAJMAL INSTITUTE OF TECHNOLOGY

Summer/Industrial Training Evaluation Form F05(MSIT-EXM-PA-02)

(Year 2014-2018)

Details of the Student

Name: Pranav Gupta

Roll No: 03196303114

Branch and Semester: IT, 5th sem

Mobile No: 9013289231

Email ID: pranav.gupta.cse@gmail.com

Details of the Organisation

Name and address of organization: Landmark Insurance Brokers Pvt Ltd, Kurls(W), Mumbai

Broader Area: General Insurance

Name of Instructor: Mr. Rajendra Raval

Designation and Contact No: General Manager-MIS 9375166823

Student_Performance_Record

	No. of days scheduled for the training	Number of days actually attended	Curriculum Scheduled for the student	Curriculum actually covered by the student
Week 1	5	5	Introduction to Organization and study of Customer Database	Same as scheduled
Week 2	5	5	Microsoft excel basics to manage data	Same as scheduled
Week 3	5	5	Making Pivot tables	Same as scheduled
Week 4	5	5	Using pivot tables to make charts	Same as scheduled
Week 5	5	5	Mining data to uncover useful hidden patterns	Same as scheduled
Week 6	5	5	Making and submission of reports	Same as scheduled

Pranav

(Signature of the Student)

Any Comments or Suggestions for the student performance during the training....Master.......PRANAV.......has.....excellent.....work.....ethics........
.and.....I......wish......him.....all.....the......luck.......in.....life.......................

(Signature of the Instructor)

CONTENTS

DATA

OVERVIEW

Data is a set of values of qualitative or quantitative variables. Pieces of data are individual pieces of information. While the concept of data is commonly associated with scientific research, data is collected by a huge range of organizations and institutions, ranging from businesses (e.g., sales data, revenue, profits, and stock price), governments (e.g., crime rates, unemployment rates, literacy rates) and non-governmental organizations (e.g., censuses of the number of homeless people by non-profit organizations).

Data is measured, collected and reported, and analyzed, whereupon it can be visualized using graphs, images or other analysis tools. Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing. Raw data ("unprocessed data") is a collection of numbers or characters before it has been "cleaned" and corrected by researchers. Raw data needs to be corrected to remove outliers or obvious instrument or data entry errors (e.g., a thermometer reading from an outdoor Arctic location recording a tropical temperature). Data processing commonly occurs by stages, and the "processed data" from one stage may be considered the "raw data" of the next stage. Field data is raw data that is collected in an uncontrolled "in situ" environment. Experimental data is data that is generated within the context of a scientific investigation by observation and recording.

IMPORTANCE OF DATA

Data is essentially the plain facts and statistics collected during the operations of a business. They can be used to measure/record a wide range of business activities - both internal and external. While the data itself may not be very informative, it is the basis for all reporting and as such is crucial in business.

Customer data are the metrics that relate to customer interaction. It can be the number of jobs, the number of enquiries, the income received, the expenses incurred, etc. In order to know about our interactions with the customer, we need data.

The importance of data cannot be under-stated as it provides the basis for reporting the information required in business operations.

DATA VS INFORMATION

An important distinction to make is the difference between Data and Information.

Data is the raw facts and statistics whereas Information is Data that is accurate and timely; specific and organized for a purpose; presented within a context that gives it meaning and relevance; and can lead to an increase in understanding and decrease in uncertainty.

Another way to look at information is as data that has been interpreted and then presented in a more meaningful context that allows a business to make decisions from.

IMPORTANCE OF INFORMATION

And this is the key importance of information - it allows a business to make informed decisions by presenting data in a way that can be interpreted by management. In this context, customer information would be useful in providing metrics surrounding client/customer engagement to determine better ways to engage or work with your clients.

However, it must be stated that the value of information lies not only in the information itself, but the actions that arise from the information. For example, if the information alerts you to poor customer satisfaction, it is only useful if this creates a change in the way the business deals with customers. Hence the information process should form part of a wider review process within the business to gain the best outcomes.

DATA, INFORMATION AND KNOWLEDGE

Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- nonoperational data, such as industry sales, forecast data, and macro economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

Information

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Knowledge

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

DATABASE

OVERVIEW

A database is an organized collection of data. It is the collection of schemas, tables, queries, reports, views, and other objects. The data are typically organized to model aspects of reality in a way that supports processes requiring information, such as modeling the availability of rooms in hotels in a way that supports finding a hotel with vacancies.

DATABASE MANAGEMENT SYSTEM

A database management system (DBMS) is a computer software application that interacts with the user, other applications, and the database itself to capture and analyze data. A general-purpose DBMS is designed to allow the definition, creation, querying, update, and administration of databases. Well-known DBMSs include MySQL, PostgreSQL, Microsoft SQL Server, Oracle, Sybase, SAP HANA, and IBM DB2. A database is not generally portable across different DBMSs, but different DBMS can interoperate by using standards such as SQL and ODBC or JDBC to allow a single application to work with more than one DBMS. Database management systems are often classified according to the database model that they support; the most popular database systems since the 1980s have all supported the relational model as represented by the SQL language.

DATA WAREHOUSE

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining.

DATA MINING

OVERVIEW

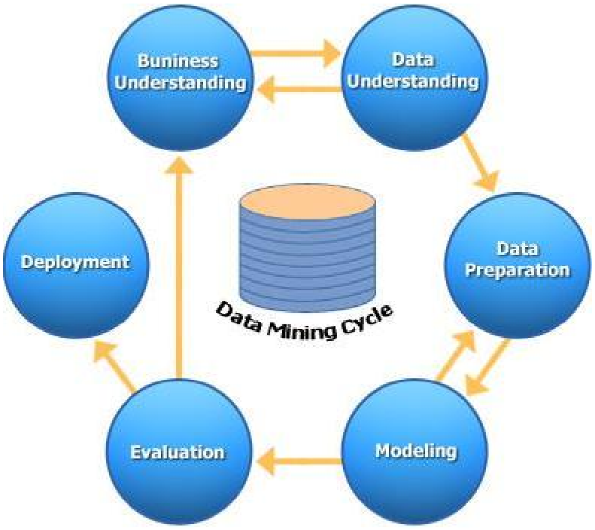
“Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.”

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step.

Data mining helps you use the discovered information in many effective ways and helps you to reduce costs and increase revenue, especially in a retail environment. By asking these types of questions and using the appropriate data, you will be able to come up with the answers:

- Who is my target customer? You need to know who is buying your product or service and tailor your marketing and sales to that demographic.
- What products are the most popular?
- What trends have been the most successful for my business?
- Should I expand my business to a new location and what is the ideal location for my business?
- How can I identify new sales prospects for my business?



Where Is Data Mining Used?

Data mining is used in a variety of fields and applications. The military use data mining to learn what roles various factors play in the accuracy of bombs. Intelligence agencies might use it to determine which of a huge quantity of intercepted communications are of interest. Security specialists might use these methods to determine whether a packet of network data constitutes a threat. Medical researchers might use them to predict the likelihood of a cancer relapse.

Some common business questions one might address through data mining methods include:

1. From a large list of prospective customers, which are most likely to respond? We can use classification techniques (logistic regression, classification trees or other methods) to identify those individuals whose demographic and other data most closely matches that of our best existing customers. Similarly, we can use prediction techniques to forecast how much individual prospects will spend.

2. Which customers are most likely to commit, for example, fraud (or might already have committed it)? We can use classification methods to identify (say) medical reimbursement applications that have a higher probability of involving fraud, and give them greater attention.

3. Which loan applicants are likely to default? We can use classification techniques to identify them (or logistic regression to assign a "probability of default" value).

4. Which customers are more likely to abandon a subscription service (telephone, magazine, etc.)? Again, we can use classification techniques to identify them (or logistic regression to assign a "probability of leaving" value). In this way, discounts or other enticements can be proffered selectively.

The Origins of Data Mining

Data mining stands at the confluence of the fields of statistics and machine learning (also known as artificial intelligence). A variety of techniques for exploring data and building models have been around for a long time in the world of statistics - linear regression , logistic regression, discriminant analysis and principal components analysis, for example. But the core tenets of classical statistics computing is difficult and data are scarce - do not apply in data mining applications where both data and computing power are plentiful.

This gives rise to Daryl Pregibon's description of data mining as "statistics at scale and speed" (Pregibon, 1999). A useful extension of this is "statistics at scale, speed, and simplicity." Simplicity in this case refers not to simplicity of algorithms, but rather to simplicity in the logic of inference. Due to the scarcity of data in the classical statistical setting, the same sample is used to make an estimate, and also to determine how reliable that estimate might be. As a result, the logic of the confidence intervals and hypothesis tests used for inference may seem elusive for many, and their limitations are not well appreciated. By contrast, the data mining paradigm of fitting a model with one sample and assessing its performance with another sample is easily understood.

Computer science has brought us “machine learning” techniques, such as trees and neural networks, that rely on computational intensity and are less structured than classical statistical models. In addition, the growing field of database management is also part of the picture.

The emphasis that classical statistics places on inference (determining whether a pattern or interesting result might have happened by chance) is missing in data mining. In comparison to statistics, data mining deals with large datasets in open-ended fashion, making it impossible to put the strict limits around the question being addressed that inference would require.

As a result, the general approach to data mining is vulnerable to the danger of “overfitting,” where a model is fit so closely to the available sample of data that it describes not merely structural characteristics of the data, but random peculiarities as well. In engineering terms, the model is fitting the noise, not just the signal.

The Rapid Growth of Data Mining

Perhaps the most important factor propelling the growth of data mining is the growth of data. The mass retailer Walmart in 2003 captured 20 million transactions per day in a 10-terabyte database (a terabyte is 1,000,000 megabytes). In 1950, the largest companies had only enough data to occupy, in electronic form, several dozen megabytes. Lyman and Varian (2003) estimate that 5 exabytes of information were produced in 2002, double what was produced in 1999 (an exabyte is one million terabytes). 40% of this was produced in the U.S.

The growth of data is driven not simply by an expanding economy and knowledge base, but by the decreasing cost and increasing availability of automatic data capture mechanisms. Not only are more events being recorded, but more information per event is captured. Scannable bar codes, point of sale (POS) devices, mouse click trails, and global positioning satellite (GPS) data are examples.

The growth of the internet has created a vast new arena for information generation. Many of the same actions that people undertake in retail shopping, exploring a library or catalog shopping have close analogs on the internet, and all can now be measured in the most minute detail.

In marketing, a shift in focus from products and services to a focus on the customer and his or her needs has created a demand for detailed data on customers.

The operational databases used to record individual transactions in support of routine business activity can handle simple queries, but are not adequate for more complex and aggregate analysis. Data from these operational databases are therefore extracted, transformed and exported to a data warehouse - a large integrated data storage facility that ties together the decision support systems of an enterprise. Smaller data marts devoted to a single subject may also be part of the system. They may include data from external sources (e.g., credit rating data).

Many of the exploratory and analytical techniques used in data mining would not be possible without today’s computational power. The constantly declining cost of data storage and retrieval has made it possible to build the facilities required to store and make available vast amounts of data. In short, the rapid and continuing improvement in computing capacity is an essential enabler of the growth of data mining.

Core Ideas in Data Mining

Classification

Classification is perhaps the most basic form of data analysis. The recipient of an offer can respond or not respond. An applicant for a loan can repay on time, repay late or declare bankruptcy. A credit card transaction can be normal or fraudulent. A packet of data traveling on a network can be benign or threatening. A bus in a fleet can be available for service or unavailable. The victim of an illness can be recovered, still ill, or deceased.

A common task in data mining is to examine data where the classification is unknown or will occur in the future, with the goal of predicting what that classification is or will be. Similar data where the classification is known are used to develop rules, which are then applied to the data with the unknown classification.

Prediction

Prediction is similar to classification, except we are trying to predict the value of a numerical variable (e.g., amount of purchase), rather than a class (e.g. purchaser or non purchaser).

Of course, in classification we are trying to predict a class, but the term “prediction” here refers to the prediction of the value of a continuous variable. (Sometimes in the data mining literature, the term “estimation” is used to refer to the prediction of the value of a continuous variable, and “prediction” may be used for both continuous and categorical data.)

Association Rules

Large databases of customer transactions lend themselves naturally to the analysis of associations among items purchased, or “what goes with what.” Association rules, or affinity analysis can then be used in a variety of ways. For example, grocery stores can use such information after a customer’s purchases have all been scanned to print discount coupons, where the items being discounted are determined by mapping the customer’s purchases onto the association rules. Online merchants such as Amazon.com and Netflix.com use these methods as the heart of a “recommender” system that suggests new purchases to customers.

Predictive Analytics

Classification, prediction, and to some extent affinity analysis, constitute the analytical methods employed in “predictive analytics.”

Data Reduction

Sensible data analysis often requires distillation of complex data into simpler data. Rather than dealing with thousands of product types, an analyst might wish to group them into a smaller number of groups. This process of consolidating a large number of variables (or cases) into a smaller set is termed data reduction.

Data Exploration

Unless our data project is very narrowly focused on answering a specific question determined in advance (in which case it has drifted more into the realm of statistical analysis than of data mining), an essential part of the job is to review and examine the data to see what messages they hold, much as a detective might survey a crime scene. Here, full understanding of the data may require a reduction in its scale or dimension to allow us to see the forest without getting lost in the trees. Similar variables (i.e. variables that supply similar information) might be aggregated into a single variable incorporating all the similar variables. Analogously, records might be aggregated into groups of similar records.

Data Visualization

Another technique for exploring data to see what information they hold is through graphical analysis. This includes looking at each variable separately as well as looking at relationships between variables. For numeric variables we use histograms and boxplots to learn about the distribution of their values, to detect outliers (extreme observations), and to find other information that is relevant to the analysis task. Similarly, for categorical variables we use bar charts and pie charts. We can also look at scatter plots of pairs of numeric variables to learn about possible relationships, the type of relationship, and again, to detect outliers.

DATA ANALYTICS

OVERVIEW

Data analytic is the science of examining raw data with the purpose of drawing conclusions about that information. Data analytics is used in many industries to allow companies and organizations to make better business decisions and in the sciences to verify or disprove existing models or theories.

Data analytics is distinguished from data mining by the scope, purpose and focus of the analysis. Data miners sort through huge data sets using sophisticated software to identify undiscovered patterns and establish hidden relationships. Data analytics focuses on inference, the process of deriving a conclusion based solely on what is already known by the researcher.

TYPES OF DATA ANALYSIS

The science is generally divided into exploratory data analysis (EDA), where new features in the data are discovered, and confirmatory data analysis (CDA), where existing hypotheses are proven true or false. Qualitative data analysis (QDA) is used in the social sciences to draw conclusions from non-numerical data like words, photographs or video.

In information technology, the term has a special meaning in the context of IT audits, when the controls for an organization's information systems, operations and processes are examined. Data analysis is used to determine whether the systems in place effectively protect data, operate efficiently and succeed in accomplishing an organization's overall goals.

SCOPE AND PURPOSE

Data analysis is the process of developing answers to questions through the examination and interpretation of data. The basic steps in the analytic process consist of identifying issues, determining the availability of suitable data, deciding on which methods are appropriate for answering the questions of interest, applying the methods and evaluating, summarizing and communicating the results.

Analytical results underscore the usefulness of data sources by shedding light on relevant issues. Some Statistics Canada programs depend on analytical output as a major data product because, for confidentiality reasons, it is not possible to release the microdata to the public. Data analysis also plays a key role in data quality assessment by pointing to data quality problems in a given survey. Analysis can thus influence future improvements to the survey process.

Data analysis is essential for understanding results from surveys, administrative sources and pilot studies; for providing information on data gaps; for designing and redesigning surveys; for planning new statistical activities; and for formulating quality objectives.

PRINCIPLES

A statistical agency is concerned with the relevance and usefulness to users of the information contained in its data. Analysis is the principal tool for obtaining information from the data.

Data from a survey can be used for descriptive or analytic studies. Descriptive studies are directed at the estimation of summary measures of a target population, for example, the average profits of owner-operated businesses in 2005 or the proportion of 2007 high school graduates who went on to higher education in the next twelve months. Analytical studies may be used to explain the behavior of and relationships among characteristics; for example, a study of risk factors for obesity in children would be analytic.

To be effective, the analyst needs to understand the relevant issues both current and those likely to emerge in the future and how to present the results to the audience. The study of background information allows the analyst to choose suitable data sources and appropriate statistical methods. Any conclusions presented in an analysis, including those that can impact public policy, must be supported by the data being analyzed.

PROCESS OF DATA ANALYSIS

Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theories.

Statistician John Tukey defined data analysis in 1961 as: "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

There are several phases that can be distinguished, described below. The phases are iterative, in that feedback from later phases may result in additional work in earlier phases.

DATA REQUIREMENTS

The data necessary as inputs to the analysis are specified based upon the requirements of those directing the analysis or customers who will use the finished product of the analysis. The general type of entity upon which the data will be collected is referred to as an experimental unit (e.g., a person or population of people). Specific variables regarding a population (e.g., age and income) may be specified and obtained. Data may be numerical or categorical (i.e., a text label for numbers).

DATA COLLECTION

Data is collected from a variety of sources. The requirements may be communicated by analysts to custodians of the data, such as information technology personnel within an organization. The data may also be collected from sensors in the environment, such as traffic cameras, satellites, recording devices, etc. It may also be obtained through interviews, downloads from online sources, or reading documentation.

DATA PROCESSING

Data initially obtained must be processed or organized for analysis. For instance, this may involve placing data into rows and columns in a table format for further analysis, such as within a spreadsheet or statistical software.

DATA CLEANING

Once processed and organized, the data may be incomplete, contain duplicates, or contain errors. The need for data cleaning will arise from problems in the way that data is entered and stored. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, identifying inaccuracy of data, and overall quality of existing data, de-duplication, and column segmentation. Such data problems can also be identified through a variety of analytical techniques. For example, with financial information, the totals for particular variables may be compared against separately published numbers believed to be reliable. Unusual amounts above or below pre-determined thresholds may also be reviewed. There are several types of data cleaning that depend on the type of data such as phone numbers, email addresses, employers etc. Quantitative data methods for outlier detection can be used to get rid of likely incorrectly entered data. Textual data spellcheckers can be used to lessen the amount of mistyped words, but it is harder to tell if the words themselves are correct.

EXPLORATORY DATA ANALYSIS

Once the data is cleaned, it can be analyzed. Analysts may apply a variety of techniques referred to as exploratory data analysis to begin understanding the messages contained in the data. The process of exploration may result in additional data cleaning or additional requests for data, so these activities may be iterative in nature. Descriptive statistics such as the average or median may be generated to help understand the data. Data visualization may also be used to examine the data in graphical format, to obtain additional insight regarding the messages within the data.

MODELING AND ALGORITHMS

Mathematical formulas or models called algorithms may be applied to the data to identify relationships among the variables, such as correlation or causation. In general terms, models may be developed to evaluate a particular variable in the data based on other variable(s) in the data, with some residual error depending on model accuracy (i.e., Data = Model + Error).

Inferential statistics includes techniques to measure relationships between particular variables. For example, regression analysis may be used to model whether a change in advertising (independent variable X) explains the variation in sales (dependent variable Y). In mathematical terms, Y (sales) is a function of X (advertising). It may be described as Y = aX + b + error, where the model is designed such that a and b minimize the error when the model predicts Y for a given range of values of X. Analysts may attempt to build models that are descriptive of the data to simplify analysis and communicate results.

DATA PRODUCT

A data product is a computer application that takes data inputs and generates outputs, feeding them back into the environment. It may be based on a model or algorithm. An example is an application that analyzes data about customer purchasing history and recommends other purchases the customer might enjoy.

COMMUNICATION

Once the data is analyzed, it may be reported in many formats to the users of the analysis to support their requirements. The users may have feedback, which results in additional analysis. As such, much of the analytical cycle is iterative.

When determining how to communicate the results, the analyst may consider data visualization techniques to help clearly and efficiently communicate the message to the audience. Data visualization uses information displays such as tables and charts to help communicate key messages contained in the data. Tables are helpful to a user who might lookup specific numbers, while charts (e.g., bar charts or line charts) may help explain the quantitative messages contained in the data.

MICROSOFT EXCEL

OVERVIEW

Microsoft Excel is a spreadsheet developed by Microsoft for Windows, Mac OS X, Android and iOS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. It has been a very widely applied spreadsheet for these platforms, especially since version 5 in 1993, and it has replaced Lotus 1-2-3 as the industry standard for spreadsheets. Excel forms part of Microsoft Office.

Microsoft Excel has the basic features of all spreadsheets, using a grid of cells arranged in numbered rows and letter-named columns to organize data manipulations like arithmetic operations. It has a battery of supplied functions to answer statistical, engineering and financial needs. In addition, it can display data as line graphs, histograms and charts, and with a very limited three-dimensional graphical display. It allows sectioning of data to view its dependencies on various factors for different perspectives (using pivot tables and the scenario manager). It has a programming aspect, Visual Basic for Applications, allowing the user to employ a wide variety of numerical methods, for example, for solving differential equations of mathematical physics, and then reporting the results back to the spreadsheet. It also has a variety of interactive features allowing user interfaces that can completely hide the spreadsheet from the user, so the spreadsheet presents itself as a so-called application, or decision support system (DSS), via a custom-designed user interface, for example, a stock analyzer, or in general, as a design tool that asks the user questions and provides answers and reports. In a more elaborate realization, an Excel application can automatically poll external databases and measuring instruments using an update schedule, analyze the results, make a Word report or PowerPoint slide show, and e-mail these presentations on a regular basis to a list of participants.

CHARTS

Excel supports charts, graphs, or histograms generated from specified groups of cells. The generated graphic component can either be embedded within the current sheet, or added as a separate object.

These displays are dynamically updated if the content of cells changes. For example, suppose that the important design requirements are displayed visually; then, in response to a user's change in trial values for parameters, the curves describing the design change shape and their points of intersection shift, assisting the selection of the best design.

FILE FORMATS

Format	Extension	Description
Excel Workbook	.xlsx	The default Excel 2007 and later workbook format. In reality a ZIP compressed archive with a directory structure of XML text documents. Functions as the primary replacement for the former binary .xls format, although it does not support Excel macros for security reasons.
Excel Macro-enabled Workbook	.xlsm	As Excel Workbook, but with macro support.
Excel Binary Workbook	.xlsb	As Excel Macro-enabled Workbook, but storing information in binary form rather than XML documents for opening and saving documents more quickly and efficiently. Intended especially for very large documents with tens of thousands of rows, and/or several hundreds of columns.
Excel Macro-enabled Template	.xltm	A template document that forms a basis for actual workbooks, with macro support. The replacement for the old .xlt format.
Excel Add-in	.xlam	Excel add-in to add extra functionality and tools. Inherent macro support because of the file purpose.

PIVOT TABLES

OVERVIEW

In data processing, a pivot table is a data summarization tool found in data visualization programs such as spreadsheets or business intelligence software. Among other functions, a pivot table can automatically sort, count, total or give the average of the data stored in one table or spreadsheet, displaying the results in a second table showing the summarized data. Pivot tables are also useful for quickly creating unweighted cross tabulations. The user sets up and changes the summary's structure by dragging and dropping fields graphically. This "rotation" or pivoting of the summary table gives the concept its name.

IMPLEMENTATION

Pivot tables are not created automatically. For example, in Microsoft Excel one must first select the entire data in the original table and then go to the Insert tab and select "Pivot Table" (or "Pivot Chart"). The user then has the option of either inserting the pivot table into an existing sheet or creating a new sheet to house the pivot table.[6] A pivot table field list is provided to the user which lists all the column headers present in the data. For instance, if a table represents sales data of a company, it might include Date of sale, Sales person, Item sold, Color of item, Units sold, per unit price, and Total price. This makes the data more readily accessible.

Date of sale	**Sales person**	**Item sold**	**Color of item**	**Units sold**	**Per unit price**	**Total price**
10/01/13	Jones	Notebook	Black	8	25000	200000
10/02/13	Prince	Laptop	Red	4	35000	140000
10/03/13	George	Mouse	Red	6	850	5100
10/04/13	Larry	Notebook	White	10	27000	270000
10/05/13	Jones	Mouse	Black	4	700	3200

The fields that would be created will be visible on the right hand side of the worksheet. By default, the pivot table layout design will appear below this list.

Each of the fields from the list can be dragged on to this layout, which has four options:

1. Report filter
2. Column labels
3. Row labels
4. Summation values

Report filter

Report filter is used to apply a filter to an entire table.

Column labels

Column labels are used to apply a filter to one or more columns that have to be shown in the pivot table.

Row labels

Row labels are used to apply a filter to one or more rows that have to be shown in the pivot table.

Summation values

This usually takes a field that has numerical values that can be used for different types of calculations. However, using text values would also not be wrong; instead of Sum it will give a count.

PIVOT CHARTS

OVERVIEW

A pivot chart is a data analysis tool that enables one to visualize a pivot table. It is a built-in feature of Microsoft Excel and Microsoft Access. PivotChart is best type of graphs for the analysis of data. The most useful feature is the possibility of quickly changing the portion of data displayed, like a PivotTable report. It makes PivotChart ideal for presentation of data in the sales reports.

INSURANCE

OVERVIEW

General Insurance helps us protect ourselves and the things we value, such as our homes, our cars and our valuables, from the financial impact of risks, big and small – from fire, flood, storm and earthquake, to theft, car accidents, travel mishaps – and even from the costs of legal action against us. And we can choose the types of risks we wish to cover by choosing the right kind of policy with the features we need.

In general, insurance works by spreading the cost of unexpected risks among a large number of people in the same region who share similar risks. When you take out an insurance policy, you pay a monthly or annual premium. That money joins the premiums of many thousands of other policyholders and goes into a big pool of funds.

If you happen to be one of the unlucky ones affected by an unexpected calamity, perhaps through severe weather or accident, that pool of funds can be used to help you up to the limit you have selected in your policy.

If things go wrong, your insurer may either repair or replace the items that have been lost or damaged, depending on the terms of your policy. You may also have the choice of receiving a cash settlement for the amount of money agreed in your policy.

INSURANCE BROKER

An insurance broker is a specialist in insurance and risk management.

Brokers act on behalf of their clients and provide advice in the interests of their clients. Sometimes an insurance broker will act as agent of an insurer too.

A broker will help you identify your individual and/or business risks to help you decide what to insure, and how to manage those risks in other ways. An insurance broker might specialize in one specific type of insurance or industry, or they might deal with many different types.

Insurance brokers can give you technical advice that can be very useful if you need to make a claim. Brokers are aware of the terms and conditions, benefits and exclusions and costs of a wide range of competing insurance policies, so they can help you find the most appropriate cover for your own circumstances.

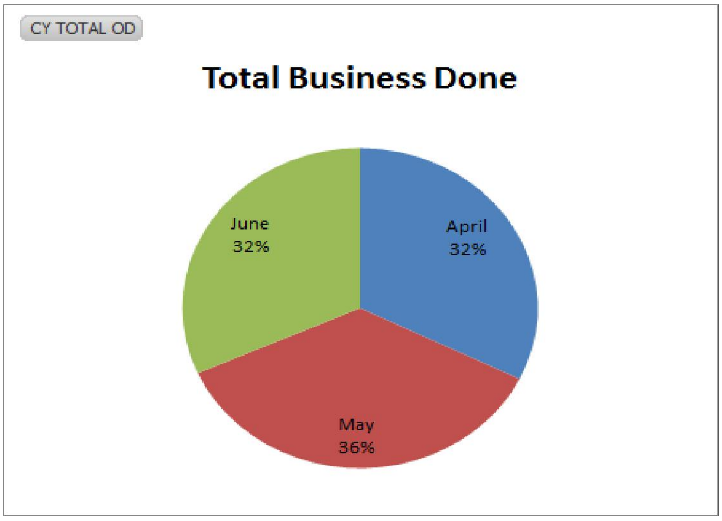
Brokers can help arrange and place the cover with the chosen insurer and can often provide advice on how to make the most of your insurance budget.

DESCRIPTION OF CUSTOMER DATABASE

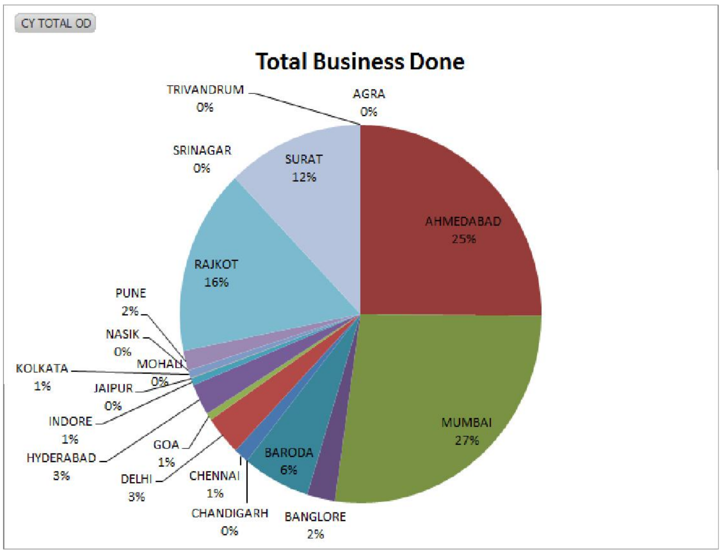
- Data set containing information regarding renewals of insurance policies.
- Data is of the first quarter (April, May, June).
- There are two types of customers: Company and Individual.
- Customers are based in various regions (cities) in India.
- There are various Insurance companies and Policy categories.
- The data contains information of a total of 19,760 customers.

REPORTS

TOTAL BUSINESS DONE

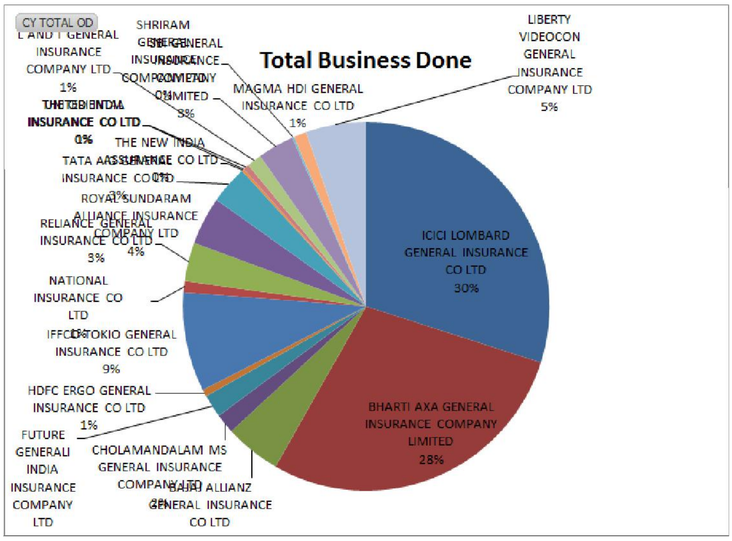


Inference
Maximum percentage of business was done in the month of May, followed by equal amounts of business in April and June.



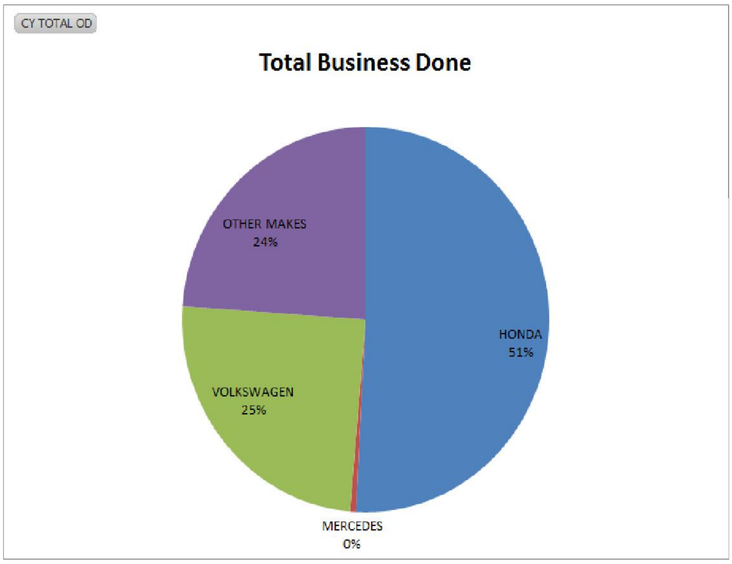
Inference

Maximum percentage of business was done in Mumbai, followed by Ahmadabad, which together constitute more than 50% of the total business.



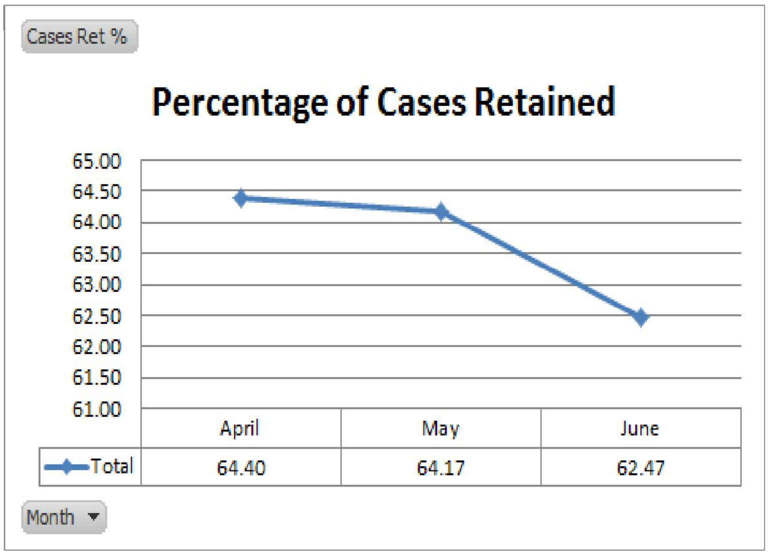
Inference

Maximum percentage of business was done with ICICI Lombard, followed by Bharti Axa, which together constitute more than 50% of the total business.

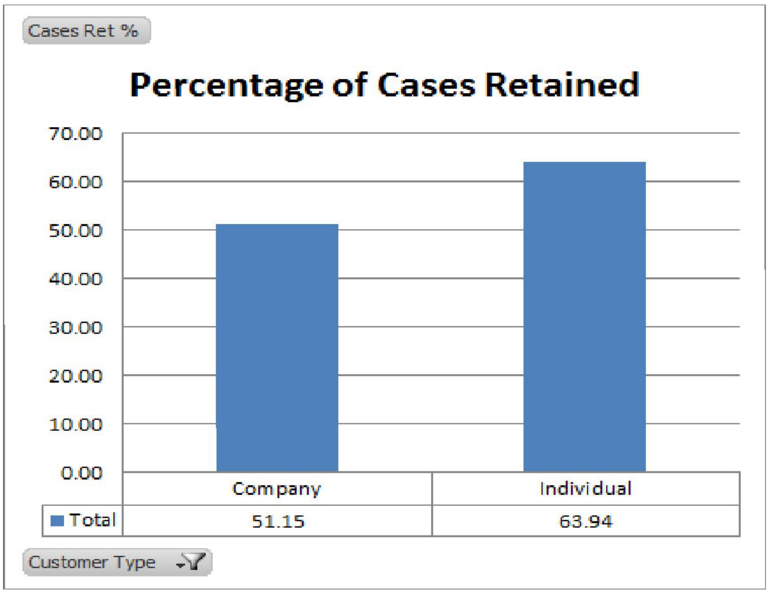


Inference
More than 50% of the business was done with vehicles made by Honda, followed by Volkwagen vehicles.

CASES RETENTION PERCENTAGE



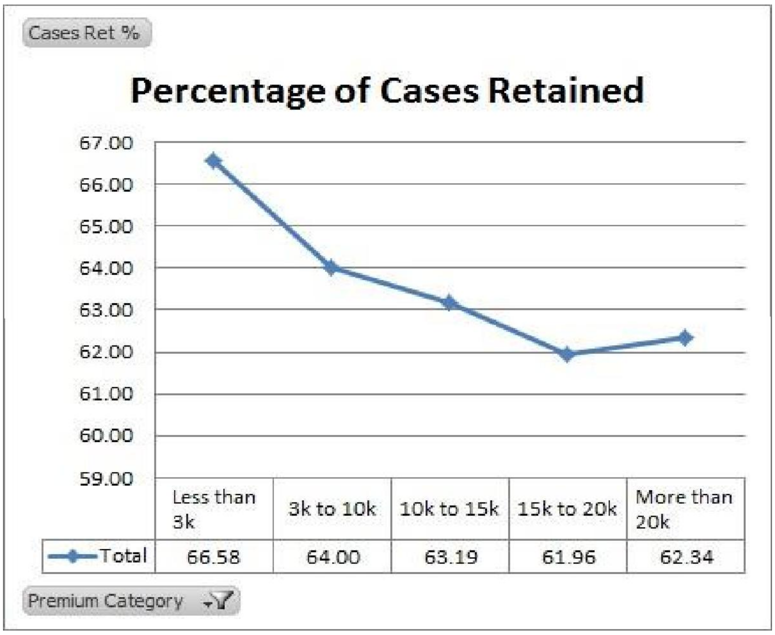
Inference
Percentage of retained cases is consistently decreasing with each month.



	Company	Individual
🟦 Total	51.15	63.94

Inference

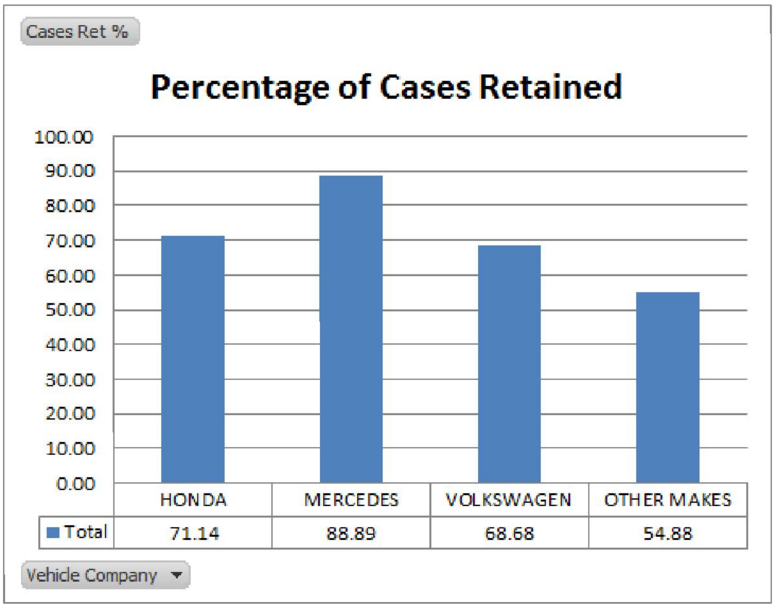
Percentage of retained cases is higher for ‘Individual’ customers than ‘Company’ customers.



	Less than 3k	3k to 10k	10k to 15k	15k to 20k	More than 20k
Total	66.58	64.00	63.19	61.96	62.34

Inference

Percentage of retained cases decreases with increase in Premium amount, except for the last 2 categories

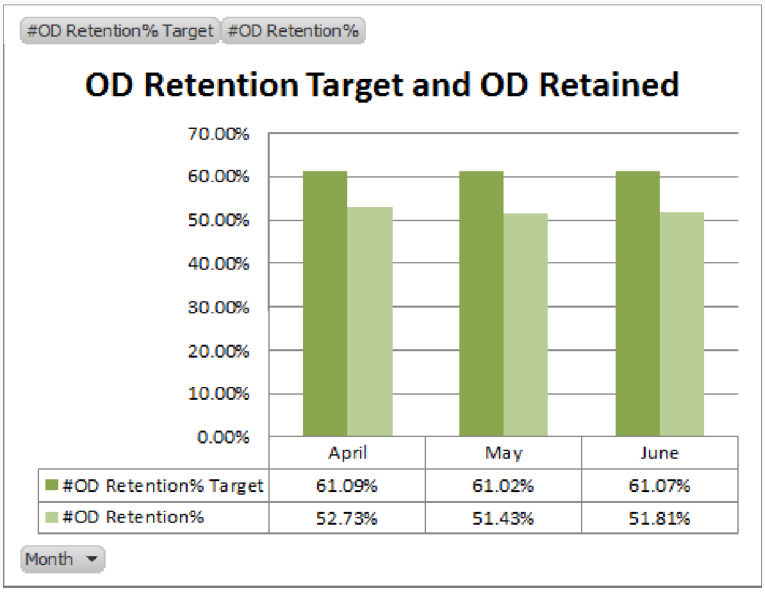


	HONDA	MERCEDES	VOLKSWAGEN	OTHER MAKES
🟦 Total	71.14	88.89	68.68	54.88

Inference

Percentage of retained cases is the highest for Mercedes vehicles.

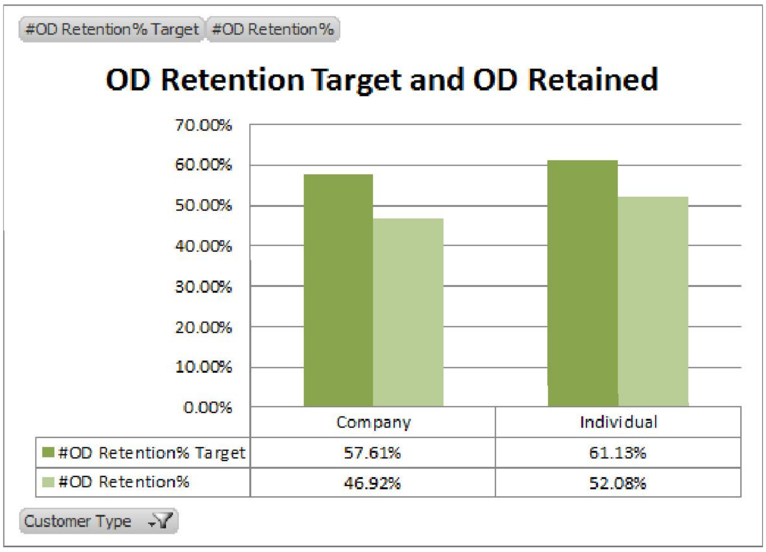
OD TARGET AND OD RETAINED



	April	May	June
🟩 #OD Retention% Target	61.09%	61.02%	61.07%
🟩 #OD Retention%	52.73%	51.43%	51.81%

Inference

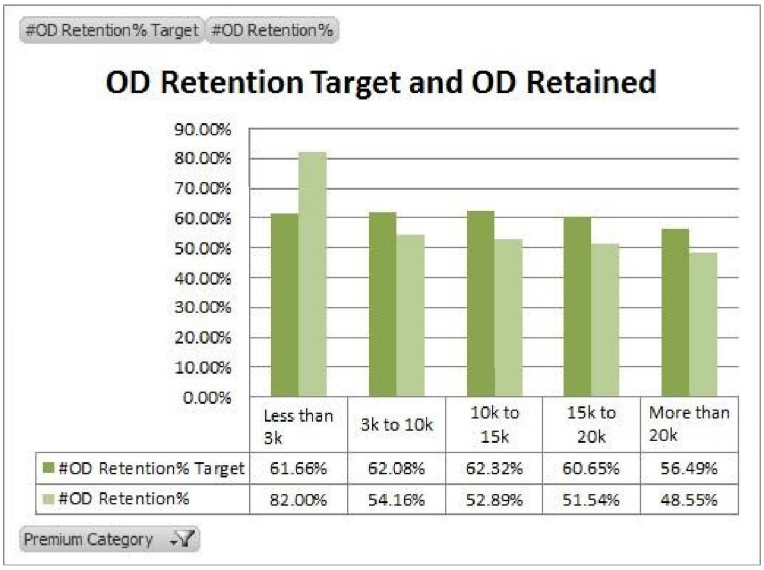
OD retention target and retained percentage is almost consistent over the three months.



	Company	Individual
🟩 #OD Retention% Target	57.61%	61.13%
🟩 #OD Retention%	46.92%	52.08%

Inference

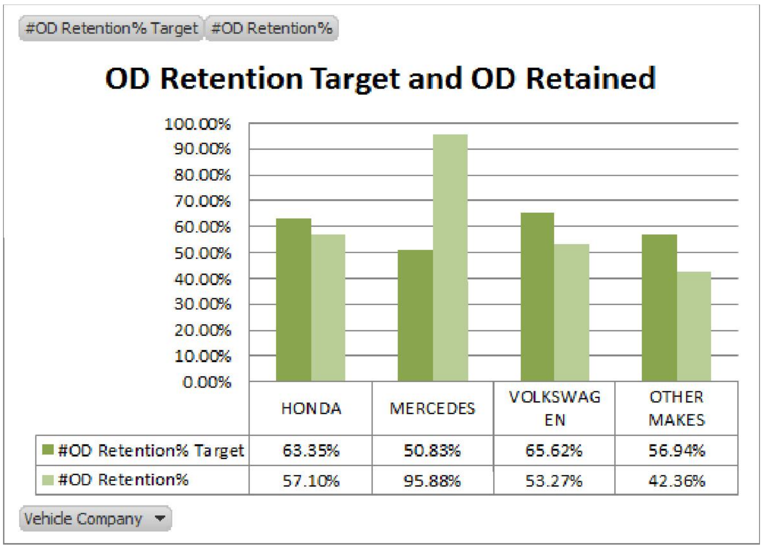
OD retention target and Retained percentage is slightly higher for ‘Individual’ customers.



	Less than 3k	3k to 10k	10k to 15k	15k to 20k	More than 20k
🟩 #OD Retention% Target	61.66%	62.08%	62.32%	60.65%	56.49%
🟩 #OD Retention%	82.00%	54.16%	52.89%	51.54%	48.55%

Inference

OD retention target is achieved in the ‘less than 3k’ category only.

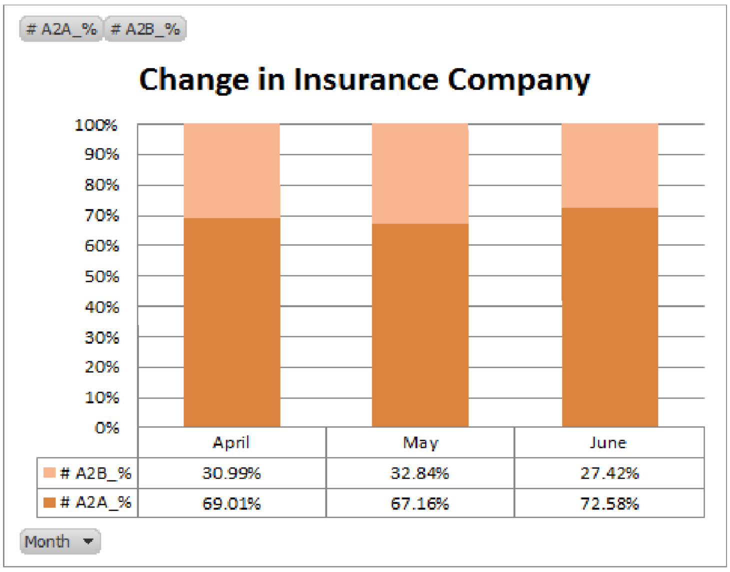


	HONDA	MERCEDES	VOLKSWAGEN	OTHER MAKES
🟩 #OD Retention% Target	63.35%	50.83%	65.62%	56.94%
🟩 #OD Retention%	57.10%	95.88%	53.27%	42.36%

Inference

OD retention target is achieved only in the case of Mercedes vehicles.

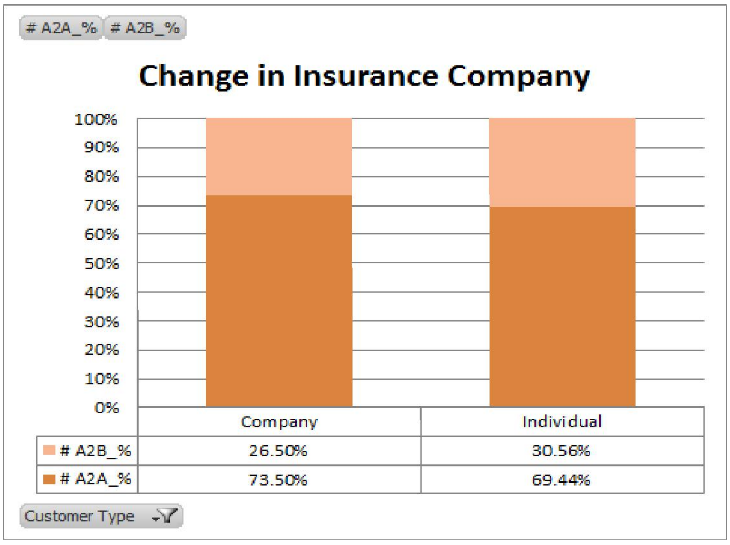
PERCENTAGE OF INSURANCE COMPANY SHIFTS



	April	May	June
🟧 # A2B_%	30.99%	32.84%	27.42%
🟧 # A2A_%	69.01%	67.16%	72.58%

Inference

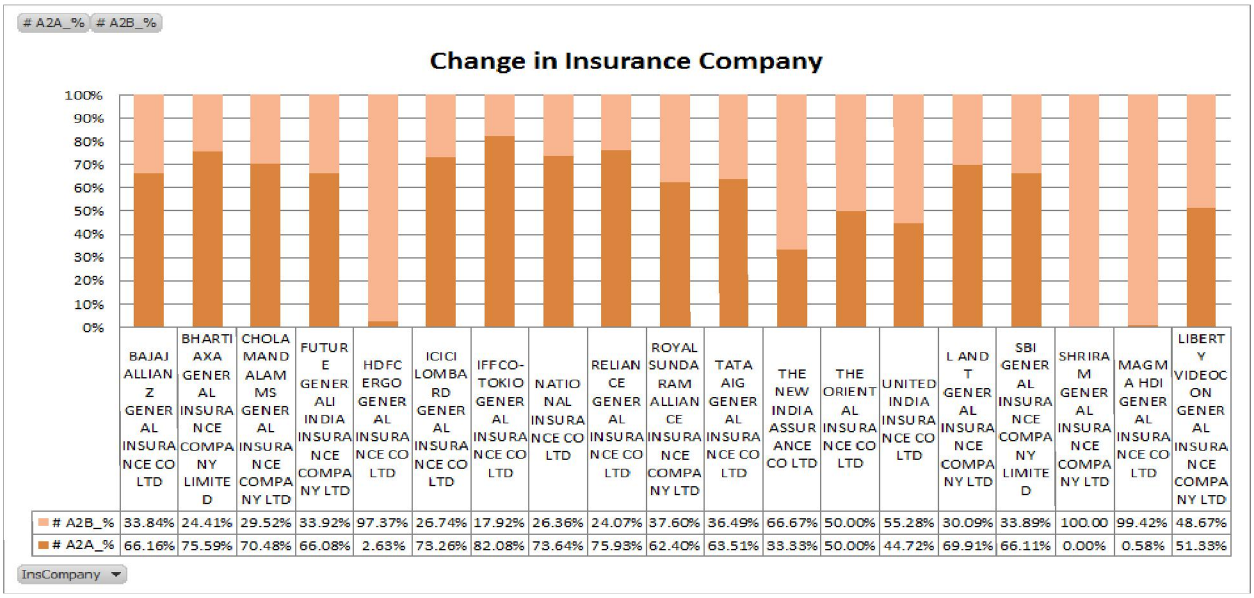
Maximum percentage of COMPANY shifts was observed in the month of May.



	Company	Individual
# A2B_%	26.50%	30.56%
# A2A_%	73.50%	69.44%

Inference

Percentage of Company shifts is higher for ‘Individual’ customers than that for ‘Company’ customers.

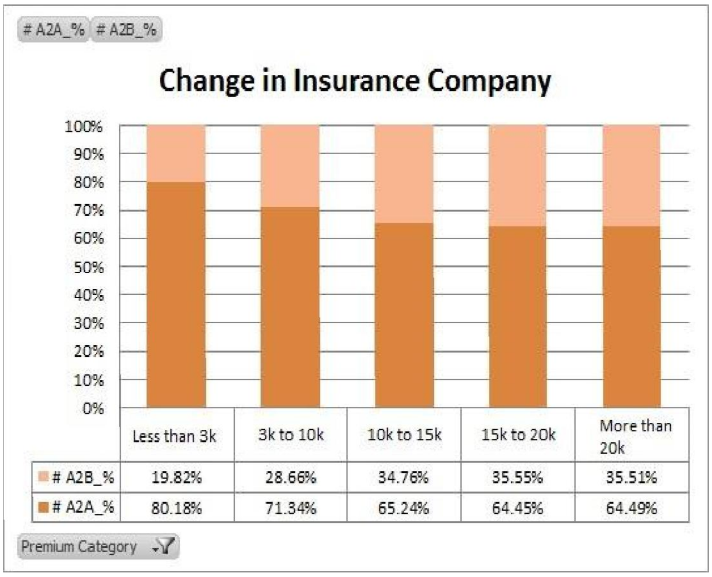


	BAJAJ ALLIANZ GENERAL INSURANCE CO LTD	BHARTI AXA GENERAL INSURANCE COMPANY LIMITED	CHOLAMANDALAM MS GENERAL INSURANCE COMPANY LTD	FUTURE GENERALI INDIA INSURANCE COMPANY LTD	HDFC ERGO GENERAL INSURANCE CO LTD	ICICI LOMBARD GENERAL INSURANCE CO LTD	IFFCO-TOKIO GENERAL INSURANCE CO LTD	NATIONAL INSURANCE CO LTD	RELIANCE GENERAL INSURANCE CO LTD	ROYAL SUNDARAM ALLIANCE INSURANCE COMPANY LTD	TATA AIG GENERAL INSURANCE CO LTD	THE NEW INDIA ASSURANCE CO LTD	THE ORIENTAL INSURANCE CO LTD	UNITED INDIA INSURANCE CO LTD	L AND T GENERAL INSURANCE COMPANY LTD	SBI GENERAL INSURANCE COMPANY LIMITED	SHRIRAM GENERAL INSURANCE COMPANY LTD	MAGMA HDI GENERAL INSURANCE CO LTD	LIBERTY VIDEOCON GENERAL INSURANCE COMPANY LTD
# A2B_%	33.84%	24.41%	29.52%	33.92%	97.37%	26.74%	17.92%	26.36%	24.07%	37.60%	36.49%	66.67%	50.00%	55.28%	30.09%	33.89%	100.00	99.42%	48.67%
# A2A_%	66.16%	75.59%	70.48%	66.08%	2.63%	73.26%	82.08%	73.64%	75.93%	62.40%	63.51%	33.33%	50.00%	44.72%	69.91%	66.11%	0.00%	0.58%	51.33%

Inference

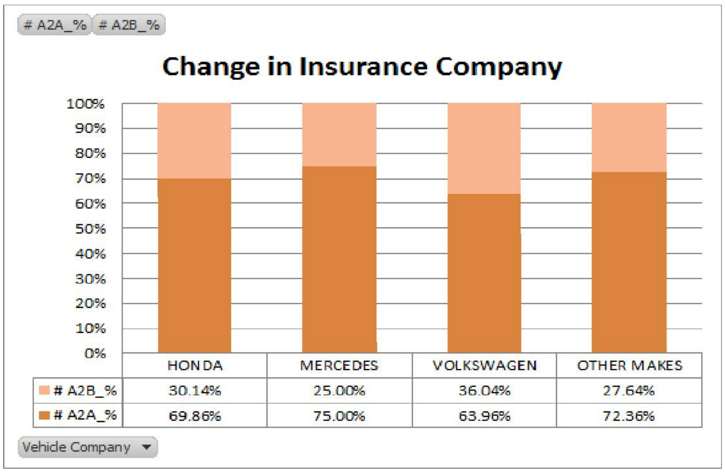
-Almost all the customers having policies of ‘Shriram Gen Ins’ and ‘Magma HDI Gen Ins’ shifted to other companies.

-Minimum percentage of shift was observed for ‘Iffco-Tokio Gen Ins’.



Inference

Percentage of shift increases consistently with increase in the Premium amount, except for the last two categories where it is almost same.



Inference

Highest shift in the Insurance Company is observed for Volkswagen vehicles, and minimum for Mercedes vehicles.