**HR ANALYTICS TOIMPROVE DECISION MAKING PROCESS**

**Section I- Introduction**

1. **Human resource analytics (HR analytics).** Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes. [[1]](#footnote-1)
2. **Machine learning**. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.[[2]](#footnote-2) Machine learning can greatly assist the HR function. By using machine learning, many traditional activities like ‘Talent Acquisition’ and ‘Employee engagement’ can be greatly improved. Machine learning can help quickly sift through thousands of job applications and shortlist candidates who have the credentials that are most likely to achieve success. This, while also helping HR managers, have access to continual insights into how their employees are feeling about their workplace and how engaged are they. Machine learning is already efficiently handle the following HR aspects[[3]](#footnote-3) :-
3. Scheduling of HR functions such as interviews, performance appraisals, group meetings and a host of other regular HR tasks.
4. Analytics and reporting on relevant HR data
5. Streamlining workflows
6. Improve recruitment procedures
7. Reducing staff-turnover
8. Personalize training
9. Measure and manage engagement
10. Enhance rewards and recognition programs
11. As the machine learning algorithm gains a deeper understanding of the company and can absorb all relevant information. In future, machine learning would be capable to do the following[[4]](#footnote-4):-
12. Identify knowledge gaps or weakness in training
13. Fine-tune and personalize training to make it more relevant and accessible to the employee
14. Become a resource for information and questions related to policies, benefits, procedures and basic conflict resolution
15. Aid in performance reviews
16. Track, guide and enhance employee growth and development
17. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly (Unsupervised Learning).
18. **Aim**. The paper aims to understand the insight into HR Analytics and the established guidelines for solving HR problems using statistics and Machine Learning Algorithms. The scope has been divided as follows:-
19. Chapter II deals with usage of **Data Analytics and Role of Data Scientists in HR Analytics.**
20. Chapter III deals with **Cross-industry standard process for data mining**, known as **CRISP-DM**,[[5]](#footnote-5) is a data mining process model that describes commonly used approaches that data mining experts use to tackle problems.
21. Chapter IV, has case study on **“Data Analytics and Machine Learning algorithms to understand reliable ways to figure out, if and why the best and most experienced employees are leaving prematurely”.**
22. In, Chapter V, **Machine Learning algorithms** have been used to solve the case study enumerated in chapter IV.
23. The use case used in the paper is a **randomly generated data set of 15000 employees generated by the author and is nowhere related to any organisation/institution.** It is merely to understanding the concepts of HR Analytics and Machine Learning advancement in the field of HR Analytics. The data set is composed of both present employees and people who have already left the organisation. [[6]](#footnote-6)
24. **Tools Used**. The programming language used for data manipulation/cleaning is ‘R’. Python has been used for coding and Visualization. SQL for querying. Also, Business Intelligence (BI) tools like Tableau have been used for generating graphs. Codes for Python used by the author for the use case has been provided as motivation for personnel.
25. **Limitations of the Paper**. The use case used in the paper is based on the randomly generated data set by the author. It is a general paper on industries/ organisations, wherein profits are measurable and accounting of HR in terms of CTC (Cost to Company) is measurable. The analysis can’t be directly applied to Government organisations, as the dynamics/ requirements of such organisations are different.

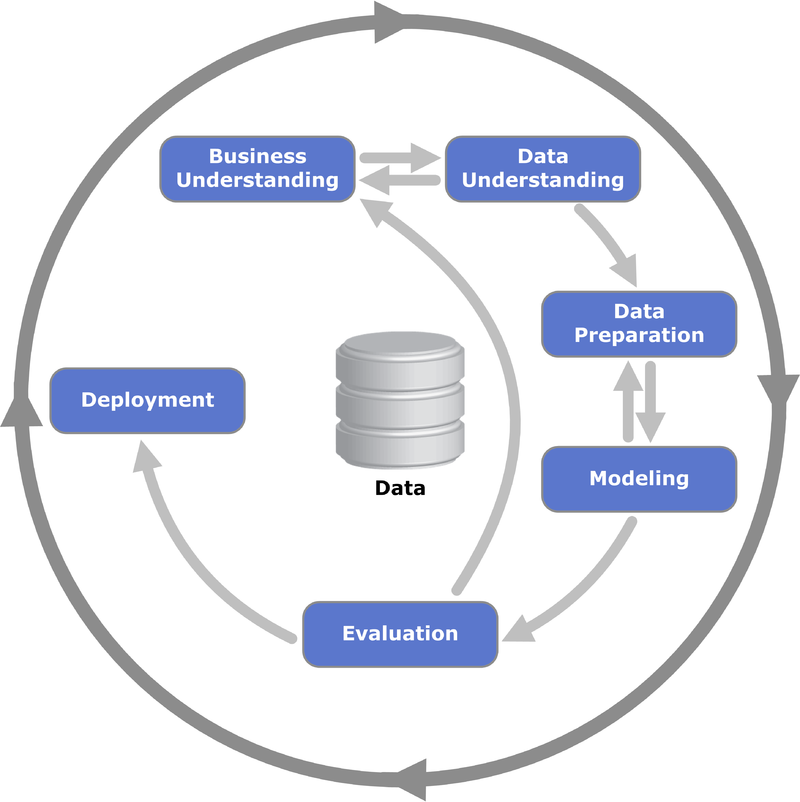
**Section II- HR Analytics Usage**

1. **Background**. One of the most important functions of human resource (HR) professionals is to evaluate talent management and development techniques and identify opportunities to more effectively manage human capital. As human behaviour is much more complex and much less predictable than that of machinery or other tangible assets, the optimization of human capital allocation has, historically, been a difficult undertaking.[[7]](#footnote-7) The use of HR analytics has noticed a recent rise in popularity in response to this challenge. By using data and metrics to design, evaluate, and implement new management policies, the “tried and true” method of using of experience, intuition, and guesswork to guide HR strategy is beginning to fall by the wayside[[8]](#footnote-8).
2. **Classification.** The relationship between an organization’s investment in human capital and its performance was first evaluated more than 50 years ago[[9]](#footnote-9). Only in the past two decades, however, has the application of data analytics to HR really taken off. Data analytics has been described as a merging of art and science[[10]](#footnote-10). While statistics are obviously a major component of any analytical exercise, analytics also involve a mental framework and logical understanding of the information at hand and the problems that need to be solved. In this way, analytics may be viewed as a “communications device,” bringing together information from multiple sources to provide an actionable representation of a current state and a likely future[[11]](#footnote-11). By providing an evidence-based approach to decision making, analytics is a logical method that enables technological manipulation of information to provide insight on relevant issues. Classification of different levels of analysis are as follows:-
3. **Descriptive Analysis**. Most commonly employed by organizations, descriptive analysis gathers data on past events or trends. This could include such measures as turnover rates or cost to hire a new employee.
4. **Predictive Analysis**. Predictive analysis evaluates why past trends have occurred and how they will change or continue without intervention. An example of predictive analysis would be the use of a model to increase the probability of selecting the right candidate for a job.
5. **Prescriptive Analysis**. . Prescriptive analysis, designs treatments for fixing current issues. This could entail creating a model to understand how alternative investments in employee training affect the organisation’s profits.
6. **Application of HR Analytics**. The application of HR analytics within a organisation may be a one-time effort or may coincide with a newly overhauled approach to organizational management. It is not uncommon, however, for one-time efforts to inspire more broad-reaching organizational change. It is important for organizational or HR leaders driving the incorporation of analytical methods to consider the purpose behind these efforts. According, analytics must be rooted in an understanding of the data to be used and the context under which that data were collected if any meaningful insight is to be gained[[12]](#footnote-12). This understanding will help determine the resources that are required and the form that the analysis will eventually take.
7. **HR Strategy**. HR professionals and management must develop a strategic understanding of how human capital contributes to organizational success prior to incorporating HR analytics. If the nature of the issue to be tackled using analytical tools is not explicitly defined, the likelihood of adding any value to the organization is extremely low. Before solutions are “fired at” the perceived issue, it is important to understand the potential causes behind the problem at hand[[13]](#footnote-13). Once the purposes behind analytical efforts are realized and an understanding of how human capital may contribute to organizational success is obtained, it is important to consider the data that will be used in subsequent analyses. Thanks to the increasing popularity of HR information systems, data are now regularly stored in one place which makes the gathering of information relatively fast and painless[[14]](#footnote-14). However, even if data are in one place, this does not mean those data are ready to be plugged in to your statistical analysis program of choice. Data quality must be considered in terms of both missing data and possible errors in data entry. An analysis of the data may require descriptive or inferential statistics (or both) and may involve descriptive, predictive, or prescriptive analyses.
8. When analyses begin to move beyond summaries of the current state and inspire “what-if” questions, an organization moves from descriptive to predictive analysis. These analyses may involve evaluations of correlation, regression models, or structural equation modeling techniques; however, more advanced data-driven decision making extend past these methods to experimental studies that identify how human capital inputs affect organizational performance[[15]](#footnote-15). It may be argued that HR analytics should facilitate experimentation to identify the causes of performance improvement and quantify the return on investment that such efforts may provide. This involves complex projects that begin with question formulation, specify a logical research design, organize data in a meaningful way, and use appropriate statistical modelling, including a variety of techniques requiring different levels of mathematical complexity. By measuring the overall impact or “lift” of an intervention, these results may then be applied more broadly to provide further improvement in different areas[[16]](#footnote-16).
9. In considering these recommendations about implementing an analytical approach, it is important to remember that support from the top of an organization is usually required to achieve success. Management support provides those driving analytical efforts with resources to allow such efforts to begin in the first place, and also with support when data are difficult to access or when such efforts are met with resistance (as is nearly always the case!).
10. **Intuition Versus Data**. There is a possibility of argument between HR professionals, about the balance of using intuition and data in making decisions. According, evidence-based management requires managerial decisions that are based on hard facts to avoid “dangerous half-truths and total nonsense” that result from a reliance on past experience, benchmarking, or commonly accepted beliefs[[17]](#footnote-17). However, it can be argued that not all decisions should be wholly grounded in analytics and that instinct and anecdote should be used in decisions involving human capital, pointing to research that most people are able to make fast and accurate judgments of personality and character. There needs to be a balance of relying on numbers and trusting common sense.
11. Analytical techniques, in and of themselves, will not provide limitless rewards to those organizations who want to seek their use. It is believed that the application of HR analytics, combined with human judgment and managerial expertise, will allow better conclusions to be reached and practices to be realized than could have resulted from following the status quo of intuition and gut reaction alone.
12. **Drawbacks of HR Analytics**. The recent and dramatic rise in the popularity of HR analytics should be accompanied by a certain skepticism over the value of these efforts. There is substantial variability in the measurement maturity of organizations. Approximately 75% of HR departments do not have usable base metrics[[18]](#footnote-18). This means that, for many organizations, there is a big leap from their current state to the appropriate use of analytics.
13. **Role of Data Scientists/Analytics and HR Professionals.**  Data scientists can help mitigate the problems with initiating successful HR analytics programs, assisting the HR profession in entering a “new world of strategic analytics-driven HR”[[19]](#footnote-19). This is true in terms of both the application of content and the use of quantitative methods. Many HR professionals lack a detailed understanding of analytical approaches. This hinders their ability to have meaningful interactions with data. Similarly, many analytics experts do not understand HR. This lack of overlapping skill and expertise leads to a mismatch between what HR information systems can do and what HR departments need. This calls for a different approach to HR analytics. Data Scientists, with an understanding of both the field of HR as well as quantitative methods, should play an important role in this approach. Also problematic to the implementation of HR analytics, the most commonly used HR information system packages typically lack the statistical functionality to enable the types of analyses that are required to solve the problems at hand (e.g., longitudinal and multivariate analysis methods). There is an opportunity for data scientists to facilitate the application of quantitative methods, developed in other contexts, to the realm of HR management and development. One such example of this application is discussed below.
14. Organizations should not implement HR analytics programs because they are trendy or because their competitors use these approaches. HR analytics should be adopted because their use can drive widespread firm improvement in the present and for years to come.
15. Most of the foreign companies and private sector have already integrated the benefits of using their data to understand which customers are most likely to churn and are using that information to engage special efforts to retain the key employee. The public sector in India and government organisations still have some progress to make to reach that level of analytics.
16. The Personnel Departments generate a huge amount of data on a daily basis; wages and benefits, recruitment, leaves, departures, social conflicts, annual evaluations, career evolution, etc. Big Data combined to predictive analytics can open a tremendous potential for HR professionals and may generate huge benefits for all stakeholders in the organization: mainly managers and employees. Some non-exhaustive list of applications of predictive analytics for HR Usage are enumerated in succeeding paragraphs.
17. **Recruitment Optimization.** Recruitment is the first time the employer gets in contact with its future employee. Past datasets combined to applications received and publicly available information can contain opportunities to make the right decisions in short listing of the candidates for future and facilitate the employee’s recruitment. It may aid into the following:-
18. Detect talents and high potentials.
19. Increase hiring success rates.
20. Predict recruitment channel effectiveness.
21. Predict employer brand strength.
22. Find the right balance between contingent and fixed workforce.
23. **Employee Performance.** Efficiently applying analytics and prediction to HR data can offer new insights into current and future performance optimization opportunities at several moments of the employee life cycle:-
    1. Predict absenteeism and work accident risks.
    2. Analyze employee engagement.
24. Analyze and predict best on-boarding processes to reduce time-to performance.
25. **Employee Retention.** HR Analytics allow to understand very accurately the employees’ motivations and what makes them stay longer in the organisation or decide to leave. Based on a few data sets and, algorithmic models are able to extract powerful analysis that aid to the management to decide about the right actions and may immediately improve employee retention rates and management decision for VR(Voluntary Retirement):-
26. Detect potential leavers and take preventive action
27. Analyze employee attrition by business unit or department
28. Predict turnover evolution on the short, mid and long run

**Section III- Cross Industry Standard Process for Data Mining Framework**

**(CRISP-DM)**

1. **Analytics Problem**. Analytics problem solving involves multiple steps like data cleaning, preparation, modelling, model evaluation etc. Completing a typical analytics project may take several months, and thus it is important to have a structure for it. The structure for analytics problem solving is called the CRISP-DM framework - Cross Industry Standard Process for Data Mining.. It involves a series of steps which are not necessarily sequential in nature. But it all starts with business understanding. Business Objectives can be articulated from the broad problem statement, so that the next steps in the CRISP-DM framework can be performed effectively. There are six steps involved in CRISP-DM framework, which are as follows:-



**Process Diagram Different Phases of CRISP-DM[[20]](#footnote-20)**

**Step 1: Business understanding**

1. Understanding the business and its specific problems is of utmost importance for data analyst. The problem needs to be clearly understood and needs to be converted into a well-defined analytics problem. Only then a brilliant strategy can be laid to solve it. In summary, to understand the business problem, one has to undertake the following steps of analysis:-
2. Determine the business objectives clearly
3. Determine the goals of data analysis

**Step 2: Data Understanding**

1. After business understanding, the next step is data understanding. The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. It is important to understand the data structure (number of files, rows, columns etc.), understand how they are related to each other and whether something looks fishy–like a date column having negative values. Broadly, we are interested in the following:-
2. The type of data sets that are available for analysis.
3. The information extracted from the datasets.
4. Exploring the data (by plotting graphs and observing them).
5. Performing quality checks on the data sets.
6. To summarize, data understanding has the following steps:-
7. Collect relevant data and classify them as:
8. Structured
9. Unstructured
10. Describe datasets
11. Create data dictionary
12. Summarize data
13. Explore data by plotting graphs
14. Check data quality in terms of:
15. Completeness
16. Correctness
17. Types of error/Missing values
18. **Plotting Charts**. A critical part of data understanding is exploring the data through plotting charts. While a line chart can be used to present a time-dependent trend, bar graphs and histograms are best used for categorical and continuous data respectively. A pie chart best summarises the share of different components in an aggregate whole, while a stacked bar chart is used to compare the share and contribution of categories across different sectors. A box plot is suitable to represent the quartile, percentile and outliers values, whereas a scatter plot summarises the variation of data points across two dimensions or two parameters. A grouped bar chart is best suited to present different sub-groups among the main categories.

**Step 3: Data Preparation**

1. The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modelling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modelling tools. Across projects, data analysts spend around 50-80% of the time on data cleaning and preparation, and therefore data preparation becomes one of the most crucial steps. Data is usually spread across different files. Collating those files together and selecting the required rows and columns based on business understanding is a major step in data preparation. After collating the data set, missing values and outliers, needs to be addressed.
2. Data preparation is considered the most crucial step because the model for analysis would be built on the data sets created. If the data set is erroneous, the solution to the problem after building a model would be erroneous too-no matter how the model is being created. To summarise, data preparation is one of the most time-consuming steps of the entire analysis. It consists of the following steps:-
3. Select relevant data
4. Integrate Data- one merge file is essential
5. Clean data
6. Construct Data: Derive new features- to reduce the no of variables
7. Format Data

**Step 4: Data Modelling**

1. It is well said that, **"If you torture the data long enough, it will confess."**  
   In this phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed. Data Modelling is the heart of data analytics. One can think of a model as a black box which takes relevant data as input and gives an output you are interested in. The following points are important for Data Modelling. The Models selected should be Succinct, Mathematically sound, Efficient and Easy to use

**Step 5: Model Evaluation**

1. At this stage in the project a model has been built, that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached. In data analytics, evaluation is when you put everything you have done to litmus tests. If the results obtained from model evaluation are not satisfactory, the whole process needs to be reiterated. If the model performs well and gives accurate results, model can be implementation. Evaluation is necessary to ensure that the model is robust and effective. Finally, implementation is the natural fruition of a project life-cycle.
2. One interesting insight is that the whole process is iterative in nature. The intelligence of a model has to evolve continuously. Model evaluation is done to verify or validate that the model developed is correct and conforms to reality. After the model is built, we need to check if the model works well on the actual data and not just the data from which it was built. Multiple models can be built for a certain phenomenon, but a lot of them may be imperfect. Model evaluation helps us choose and build a perfect model. Model evaluation helps us choose the best model among a given set of possible models that can be built. Comparison should be done to assess the performance of different models in various situations and then decide the best model class or algorithm to use for the required business problem.

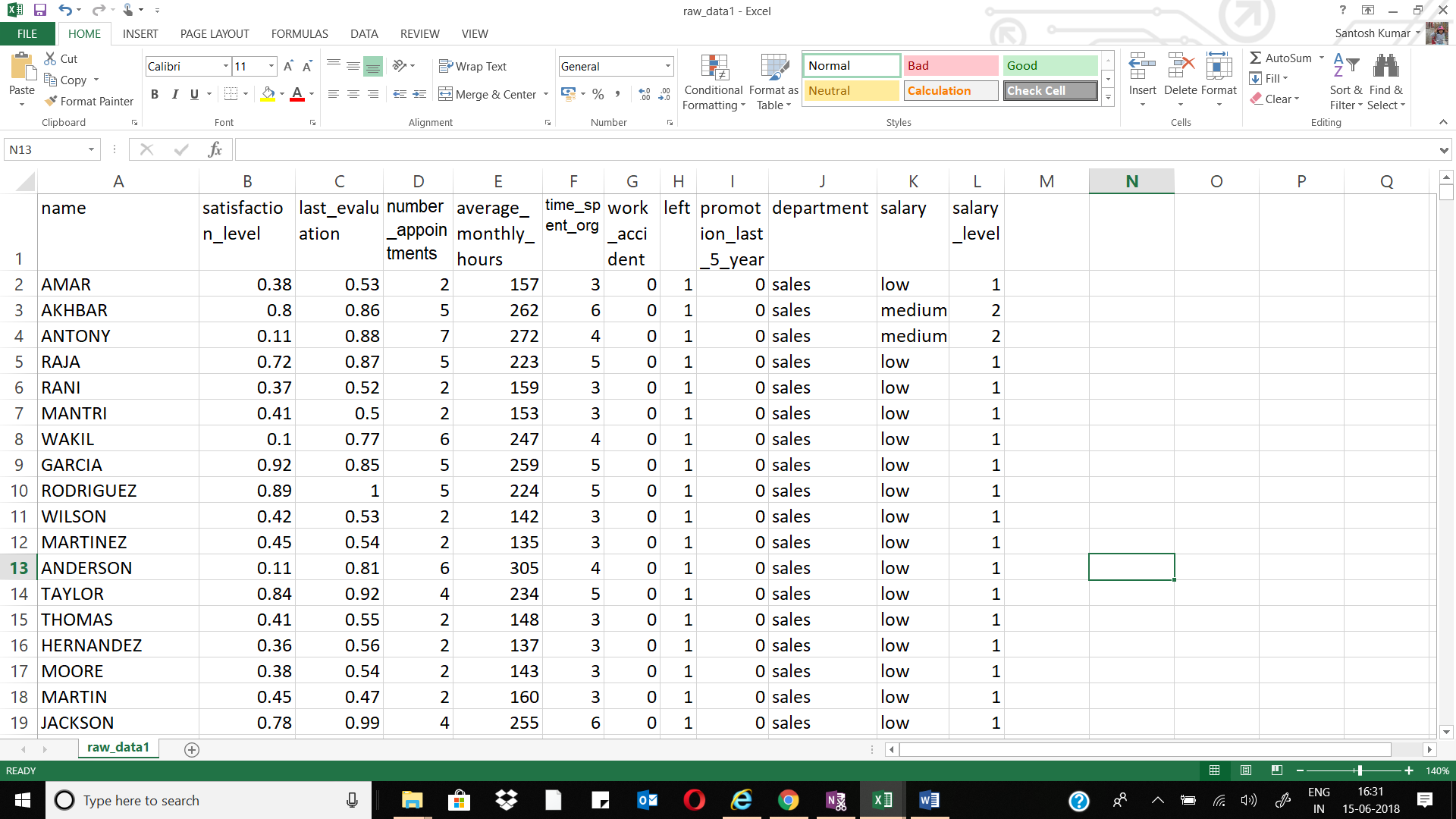
**Step 6: Model Deployment**

1. Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that is useful to the customer. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data scoring (e.g. segment allocation) or data mining process.

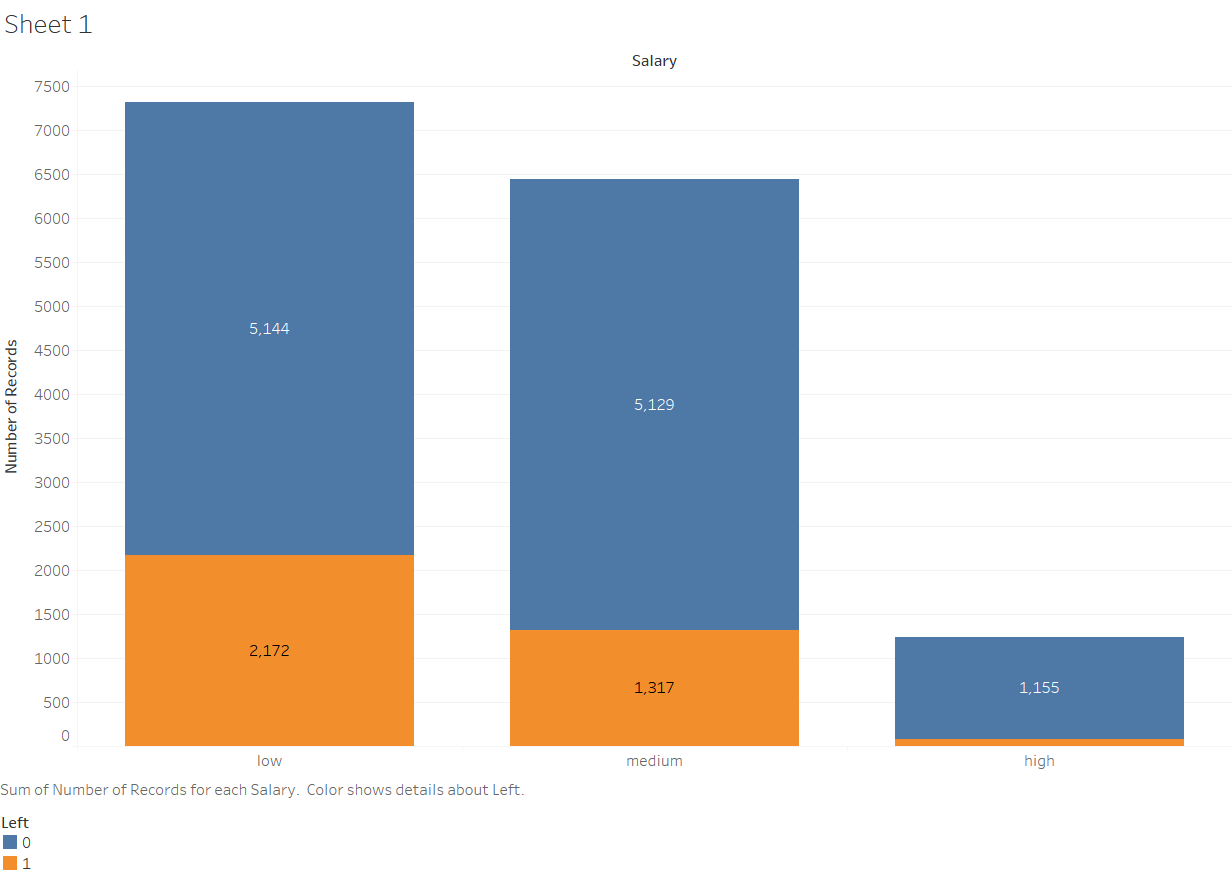
**Section IV - Use Case**

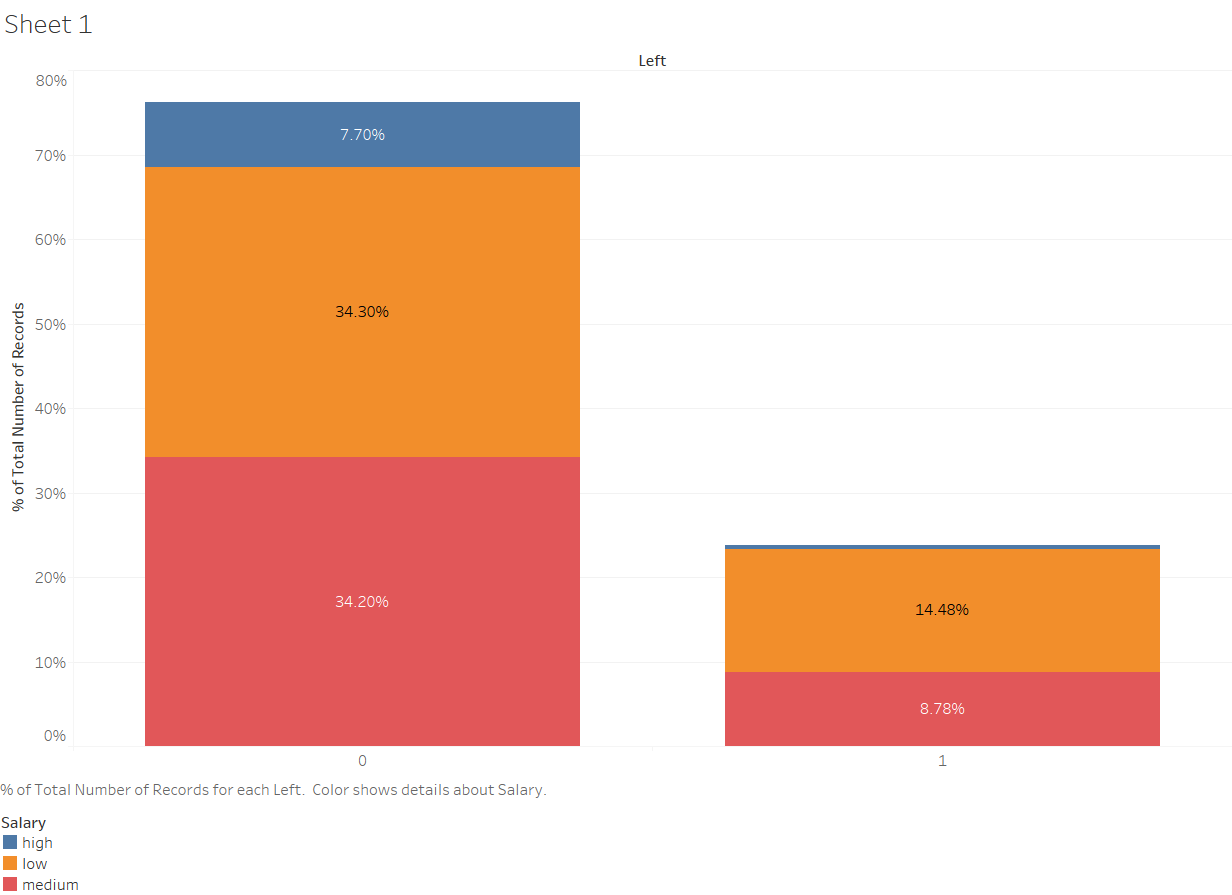
**Reduce Employee Attrition and Make Talents Stay Longer**

1. **Aim/ Business understanding.** Retaining key employees is a major stake for each organization. But are there **reliable ways to figure out if and why the best and most experienced employees are leaving prematurely**?
2. Personnel Departments generate a huge amount of data on a daily basis: wages and benefits, recruitment, leaves, departures, social conflicts, annual evaluations, career evolution, etc. Big Data combined to predictive analytics can open a tremendous potential for HR professionals and may generate huge benefits for all stakeholders in the organization.
3. The data set in the use case is a randomly generated data set of 15000 employees generated by the author and is nowhere related to any organisation/institution. It is composed of both currently working employees and people who have already left the organisation.
4. **Data Understanding/ Data Discovery.** ‘R’ programming language has been used to clean the data and to dig into the different columns of the data set. Fields in the data set include:-
5. **name**. The name of the employee.
6. **satisfaction\_level**. Employee satisfaction level. Ranges between 0 and 1. 1- indicates highly satisfied.
7. **last\_evaluation**. The grade the employee got at their last evaluation. Ranges between 0 and 1.
8. **number\_appointments**. The number of simultaneous appointments (duties) /projects the employee has worked on.
9. **average\_monthly\_hours**. The number of monthly hours the employee is working.
10. **time\_spent\_org**. The number of years the employee has been working for the company
11. **work\_accident./ Medical Issues**. Whether the employee has already had a work accident in the past (1 for yes, 0 for no)
12. **promotion\_last\_5\_years**. Whether the employee has got promoted during last 5 years. (1 for yes, 0 for no)
13. **department**. The department the employee is working,
14. **salary**. The current level of salary of the employee (3 categories : high, medium, low)
15. **left**. Whether the employee has left. 0 if the employee is still working for the company, 1 if not.



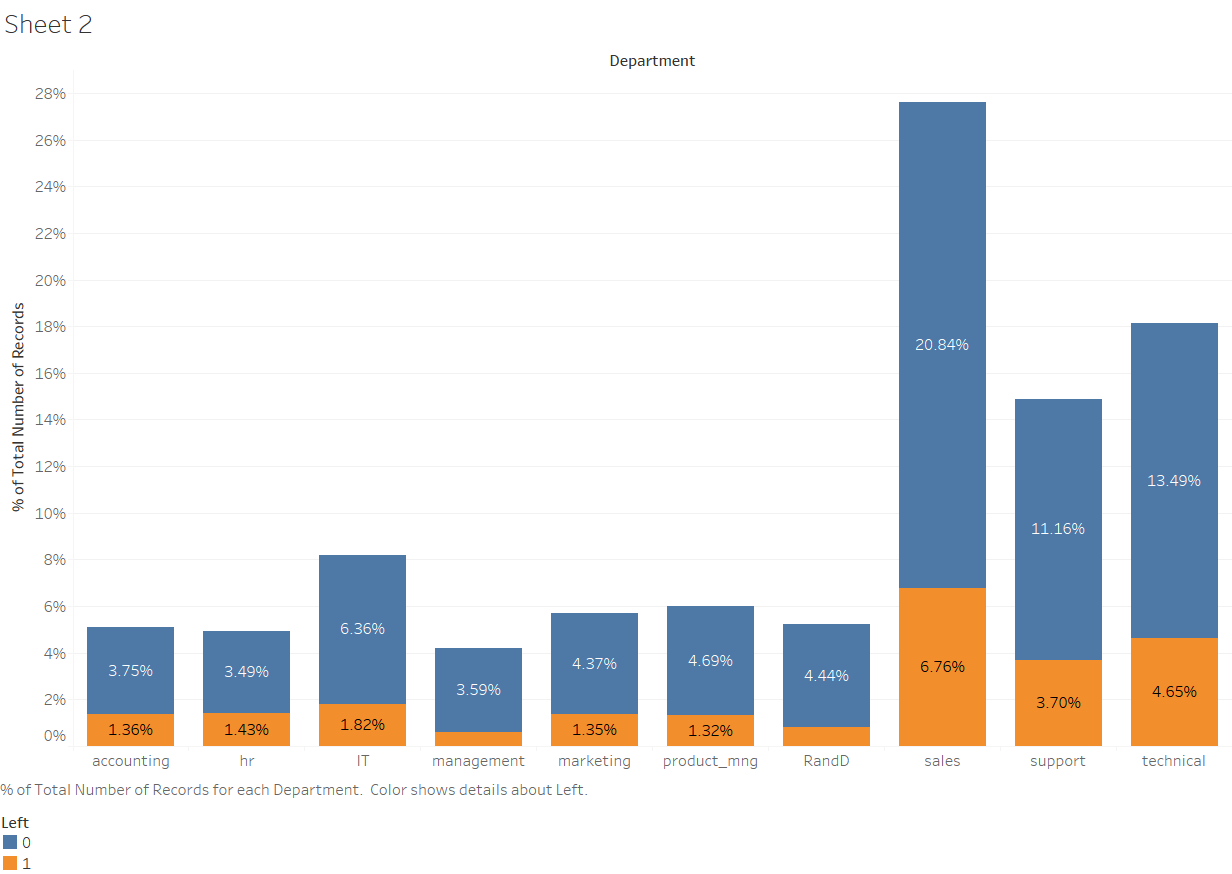
1. Prior to doing any specific analytics, it is important to understand, if how, the dataset might be correlated to each other.

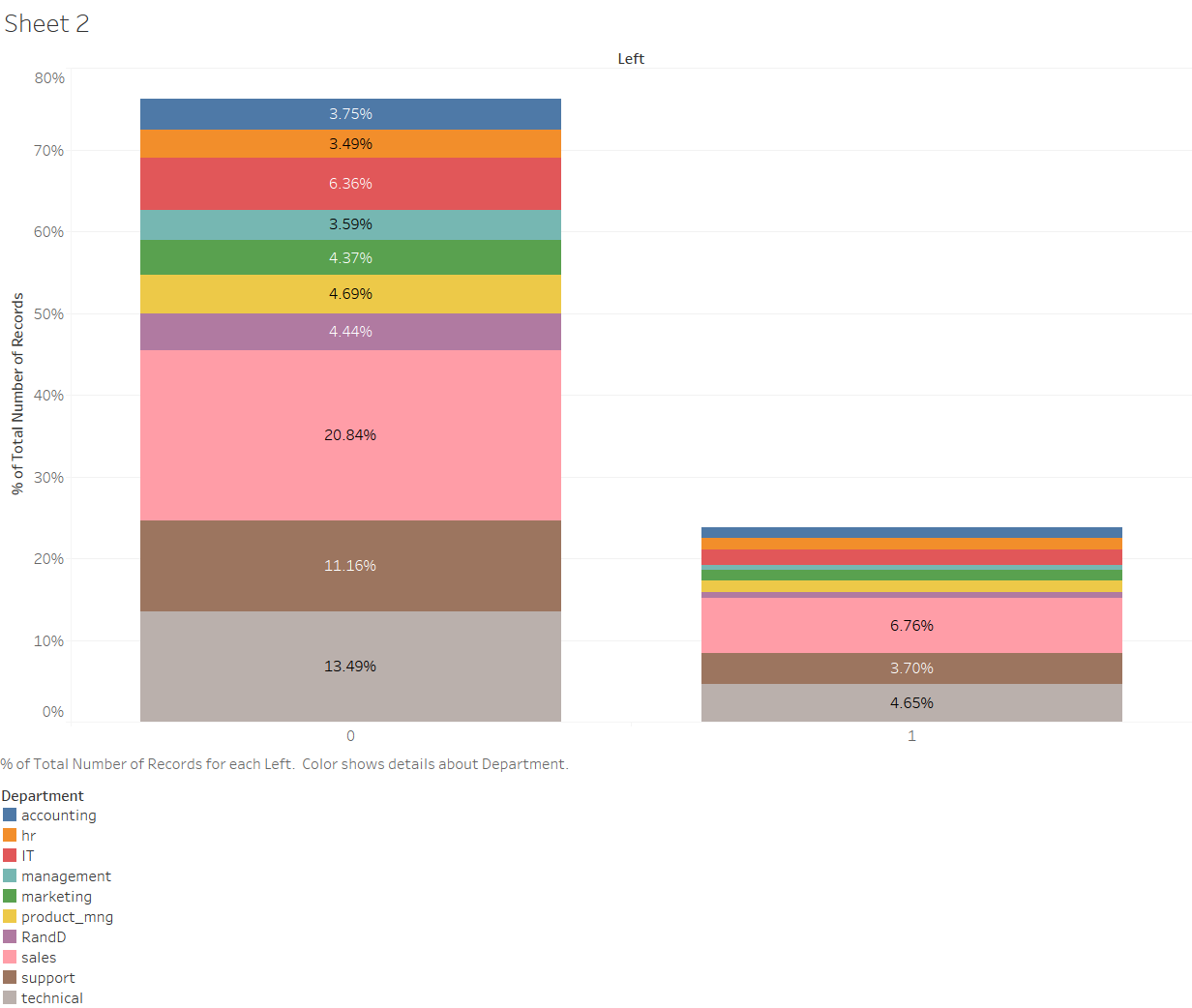




**RELATIONSHIP BETWEEN PERSONNEL LEFT(0-WORKING, 1-LEFT) AND SALARY(3- HIGH, 2-MEDIUM, 1LOW)**

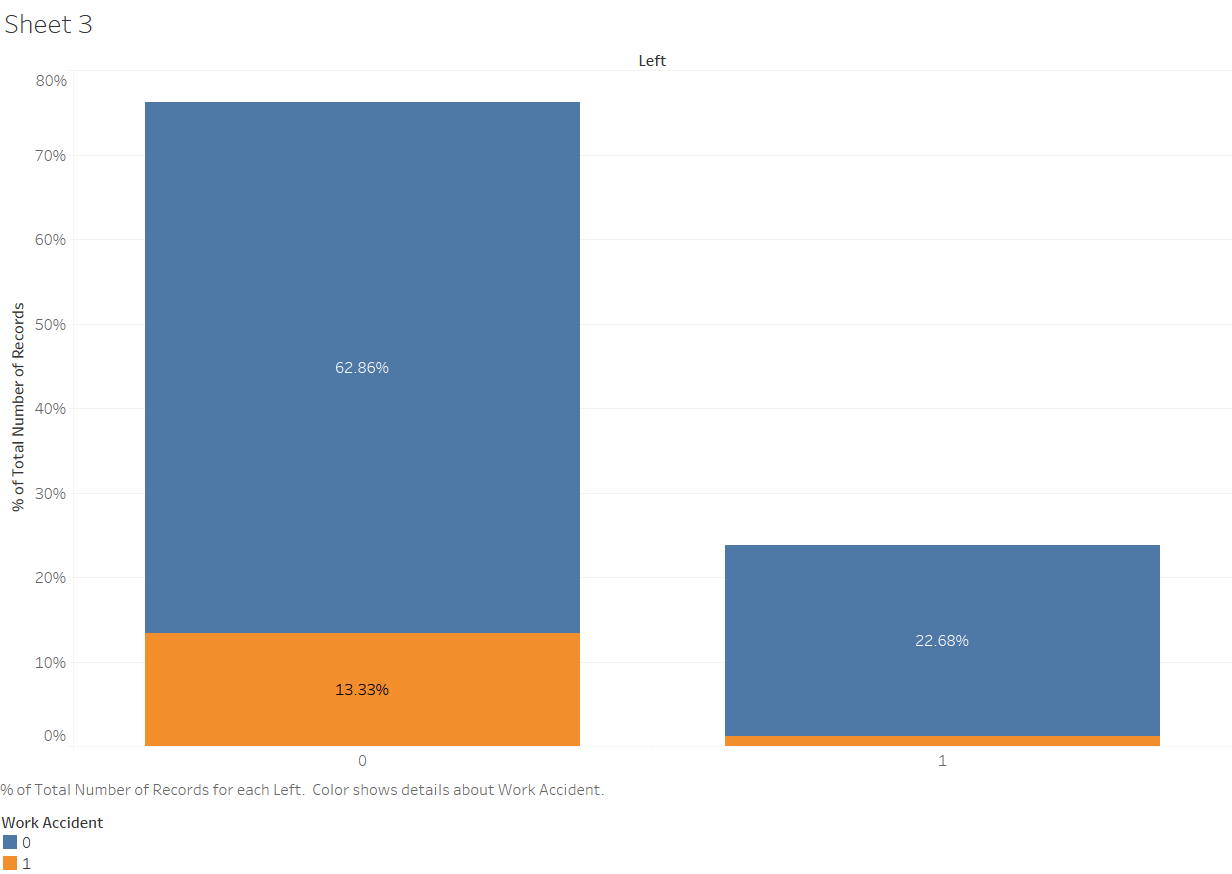
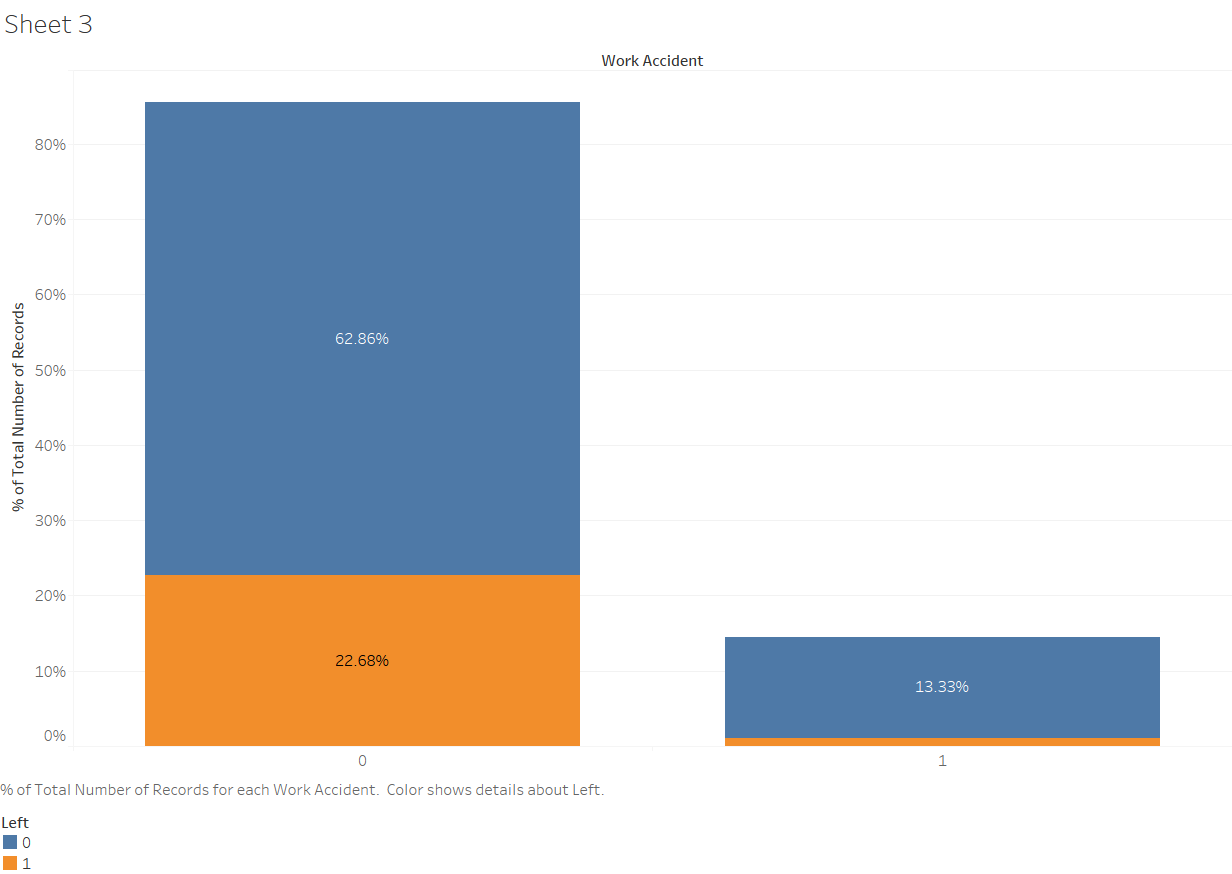
*(***Interpretation***- The number of personnel leaving are highest from the Low Salary group. It accounts for 14% of the total dataset)*





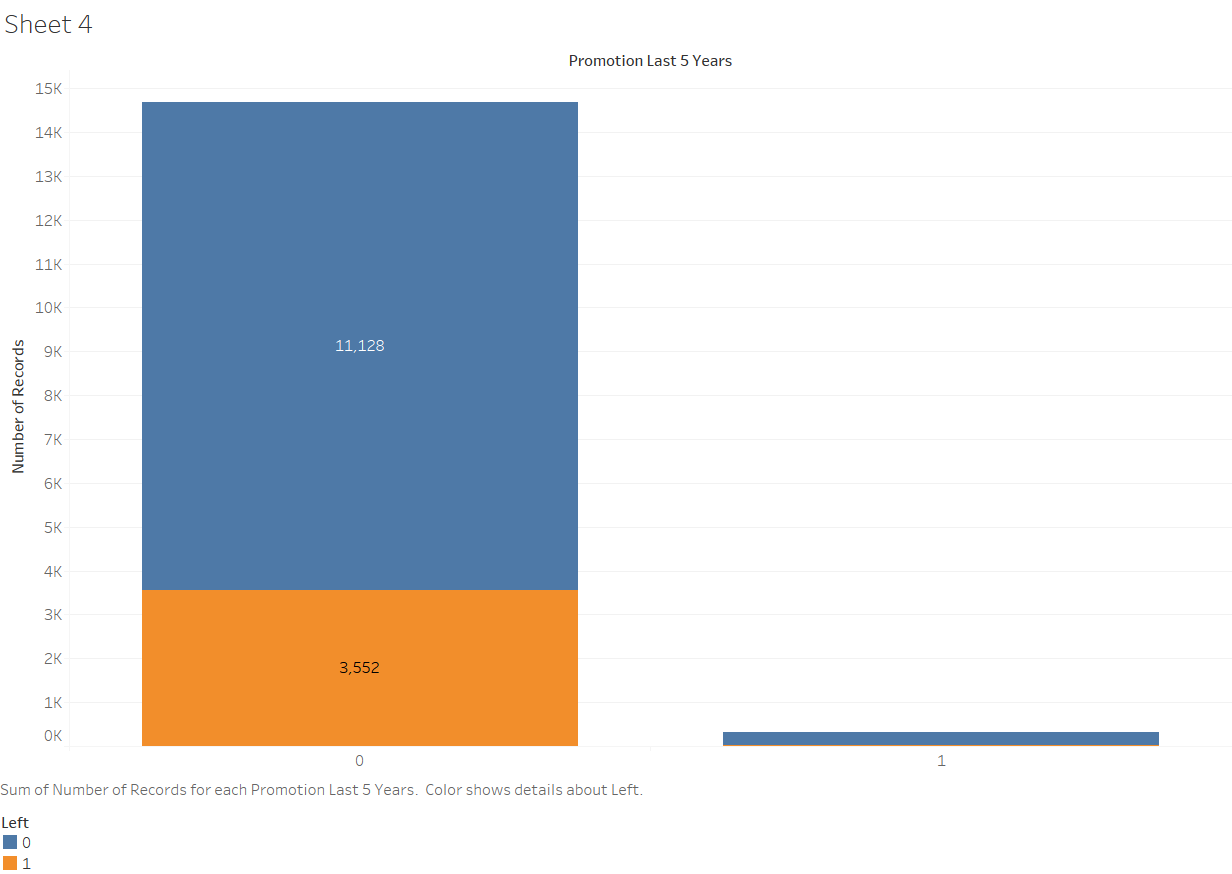
**RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND DEPARTMENTS)**

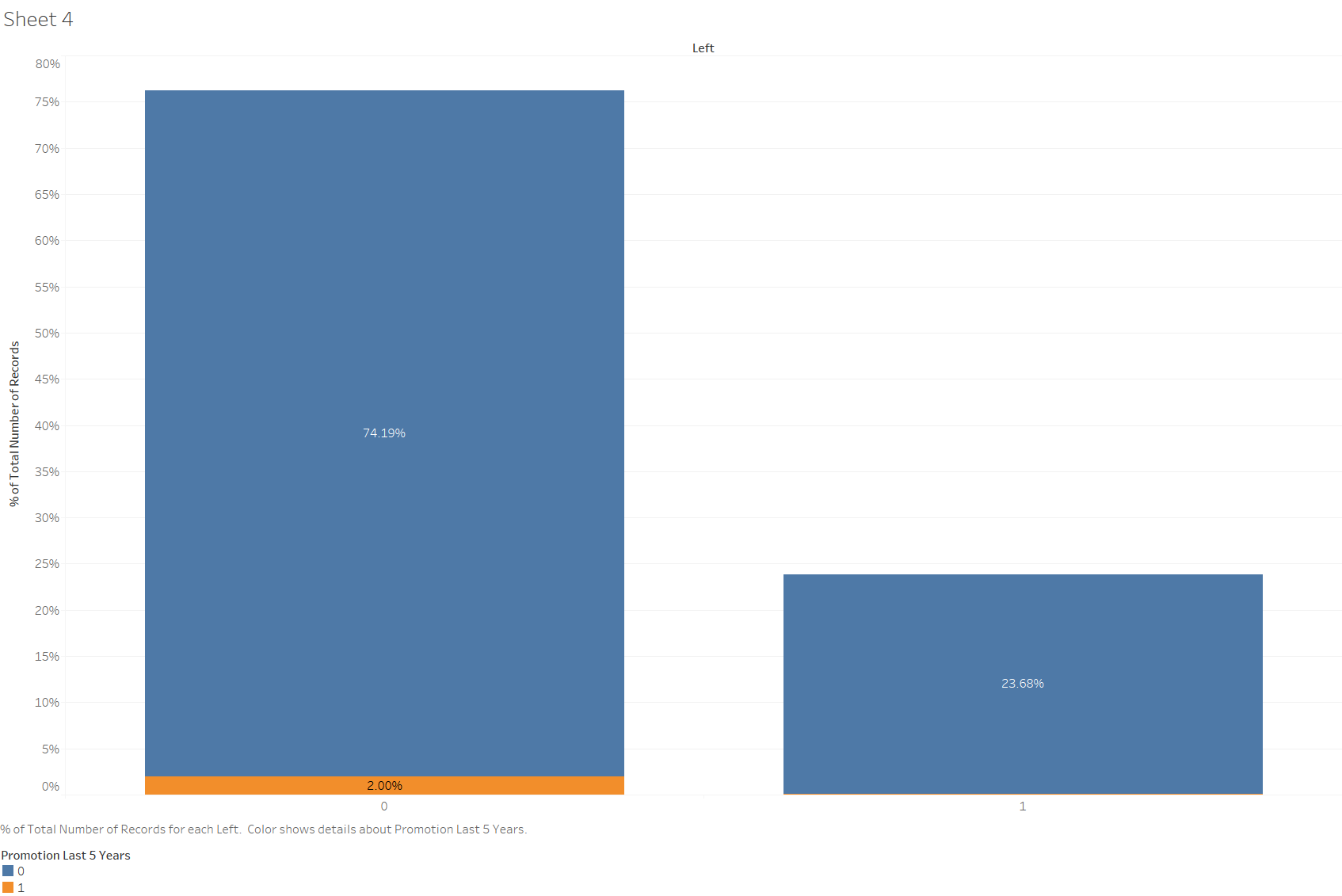
*(***Interpretation***- The number of personnel leaving are highest from Sales department followed by technical and support (% of the total dataset)*



**RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND WORK ACCIDENT (1- YES, 0- NO )**

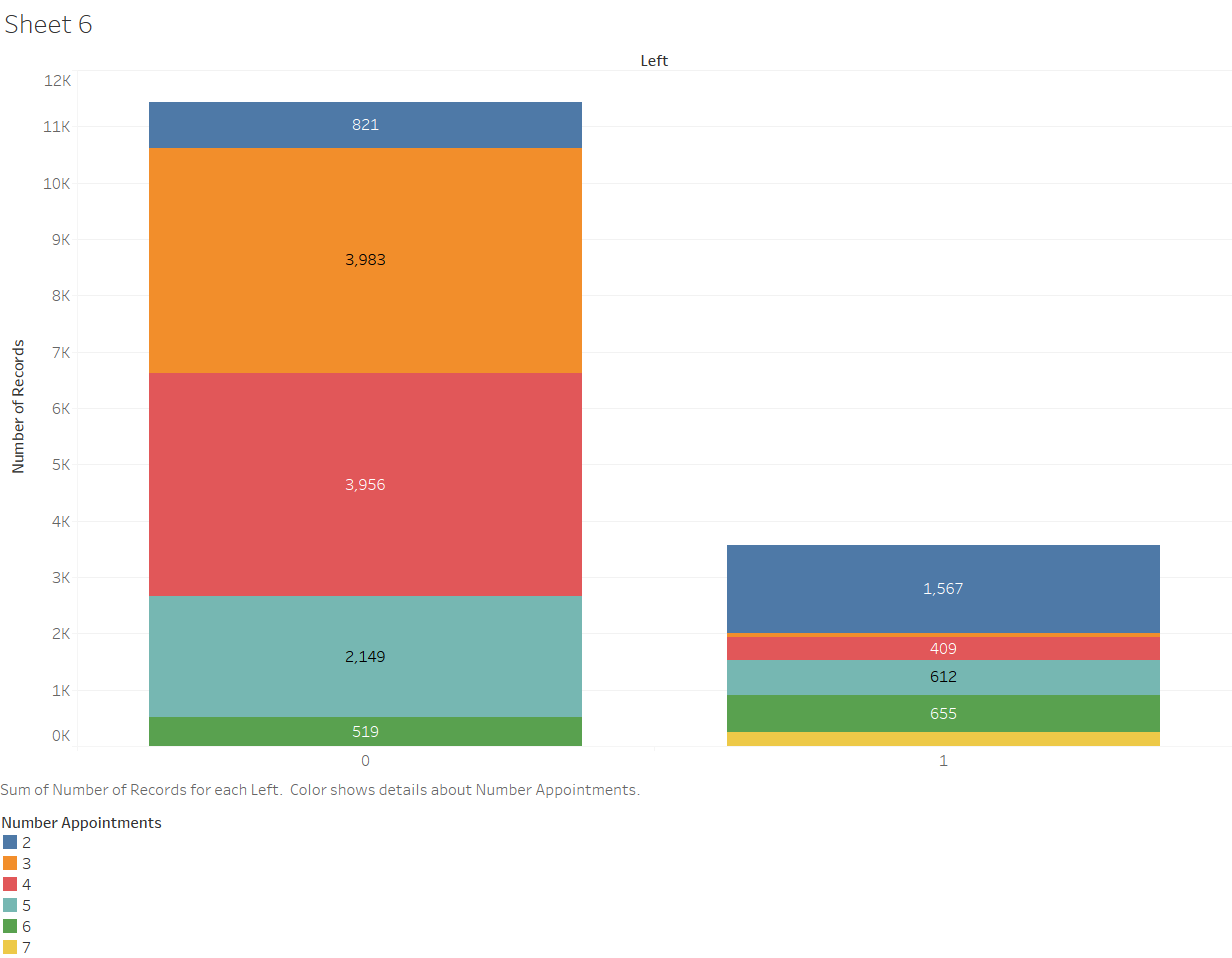
*(***Interpretation***- The number of personnel who met with any accident, normally do not leave the organisation)*

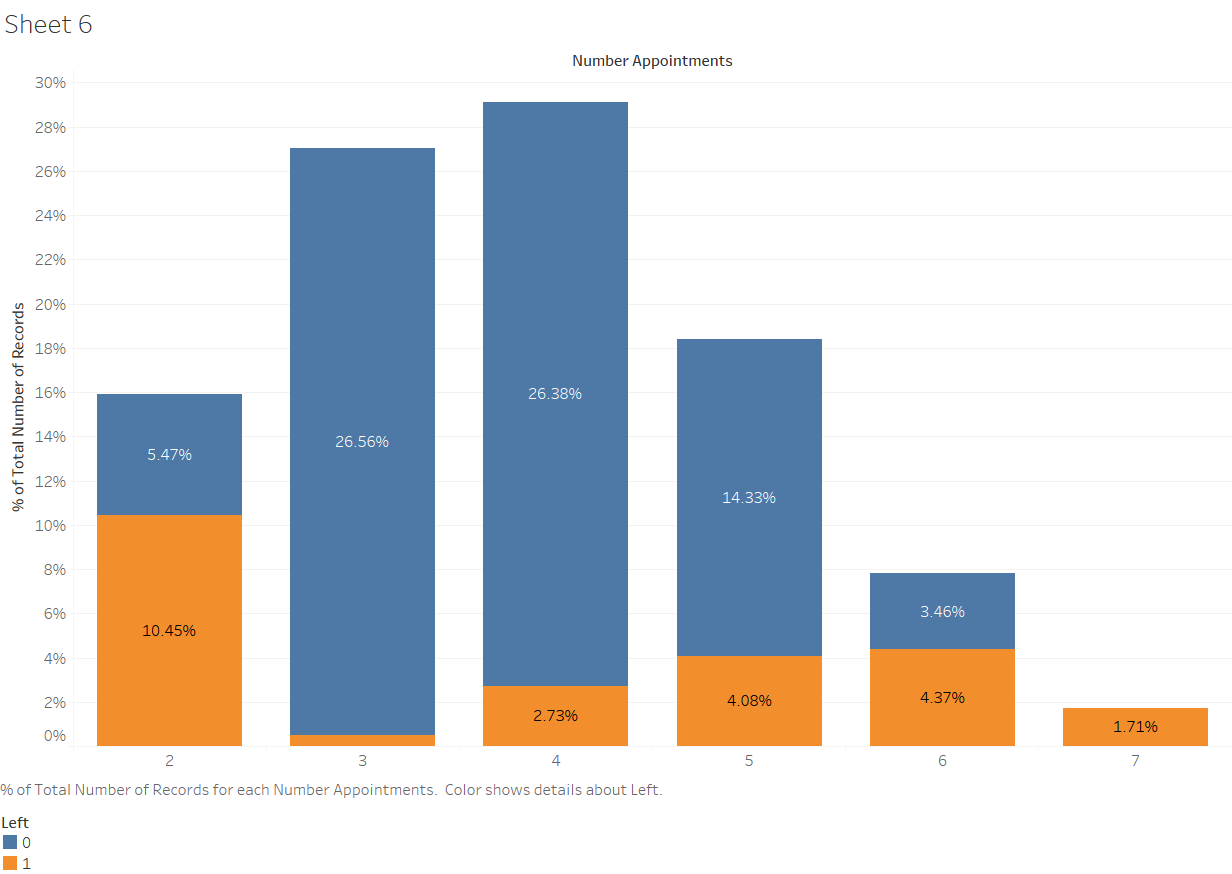




**RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND PROMOTION IN LAST FIVE YEAR (1- YES, 0- NO )**

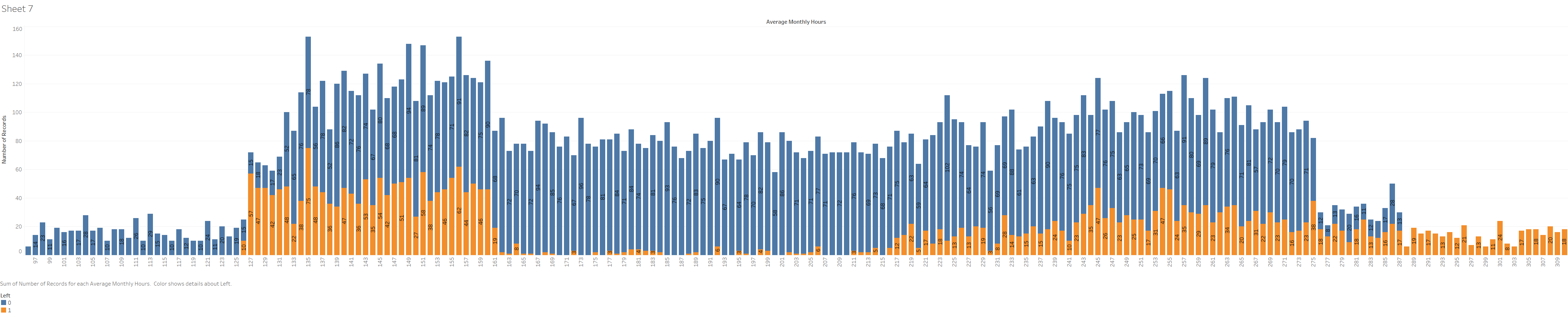
*(***Interpretation***- Personnel who have left the organisation, have not been promoted in last five years)*





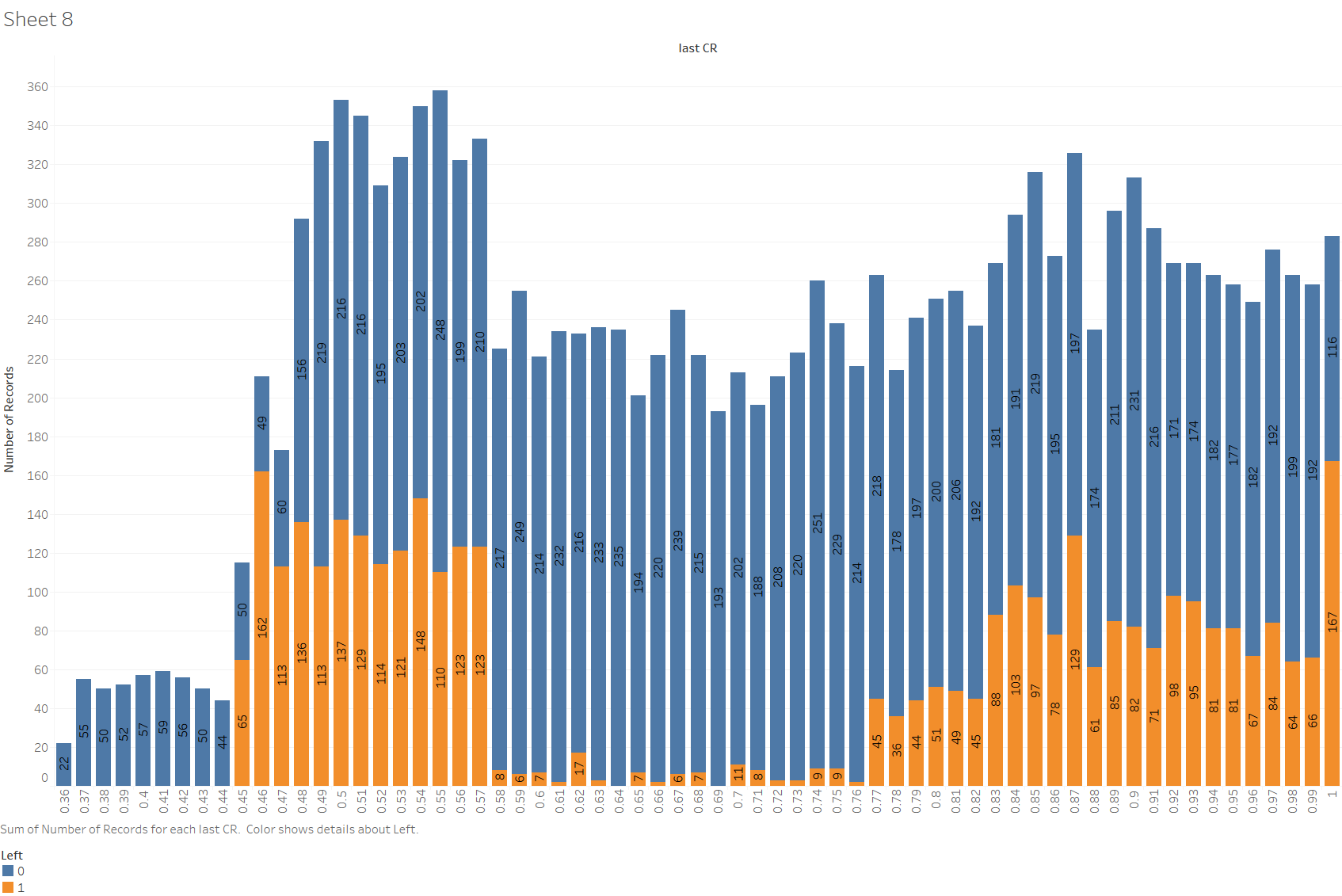
**RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT), AND NUMBER OF SIMULTANEOUS APPOINTMENTS/PROJECTS**

*(***Interpretation***- Maximum number of Personnel have left large number of of simultaneous appointments/projects(>6 .)*



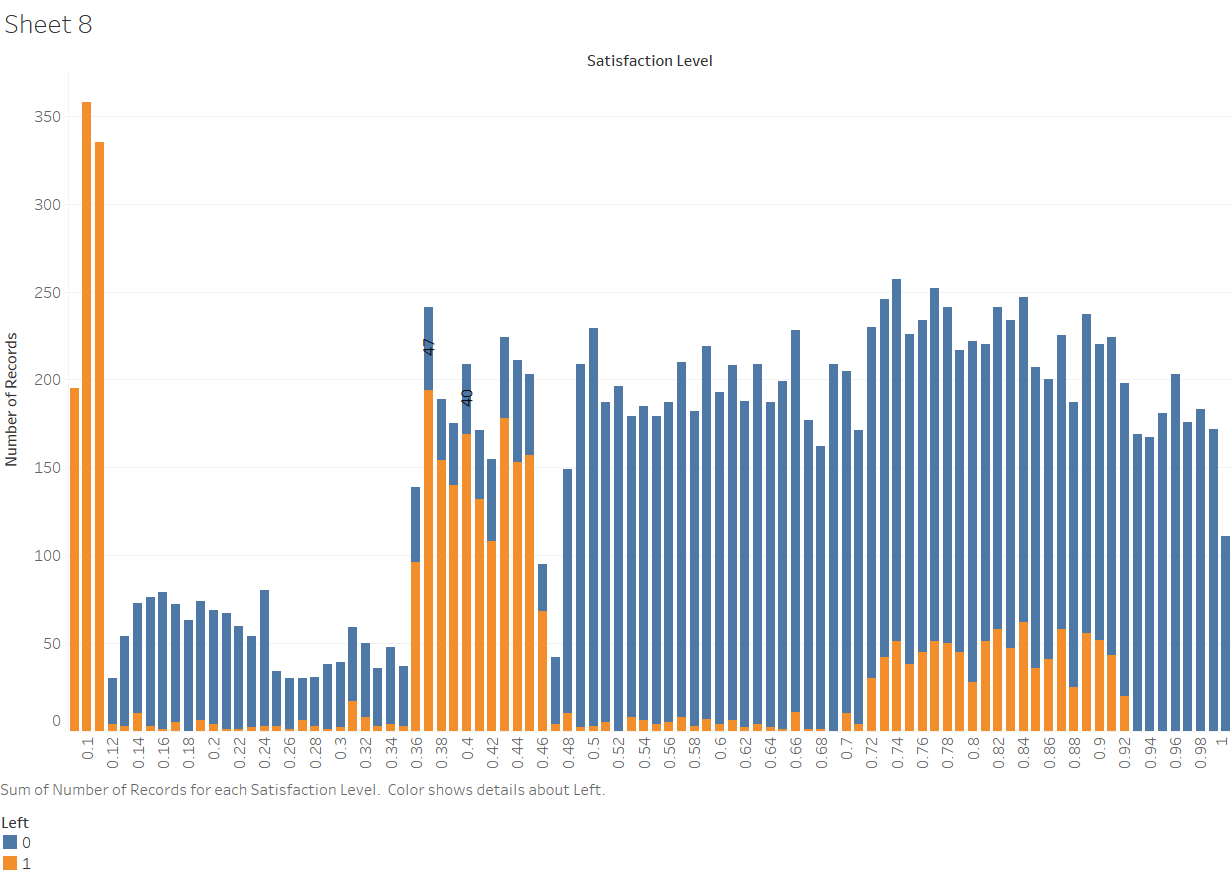
**RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND NUMBER OF WORKING HOURS/MONTH)**

*(***Interpretation***- As the number of hours/month has increased, the resignation rate has increased. All employees who have put more than 287 hrs/month have left)*



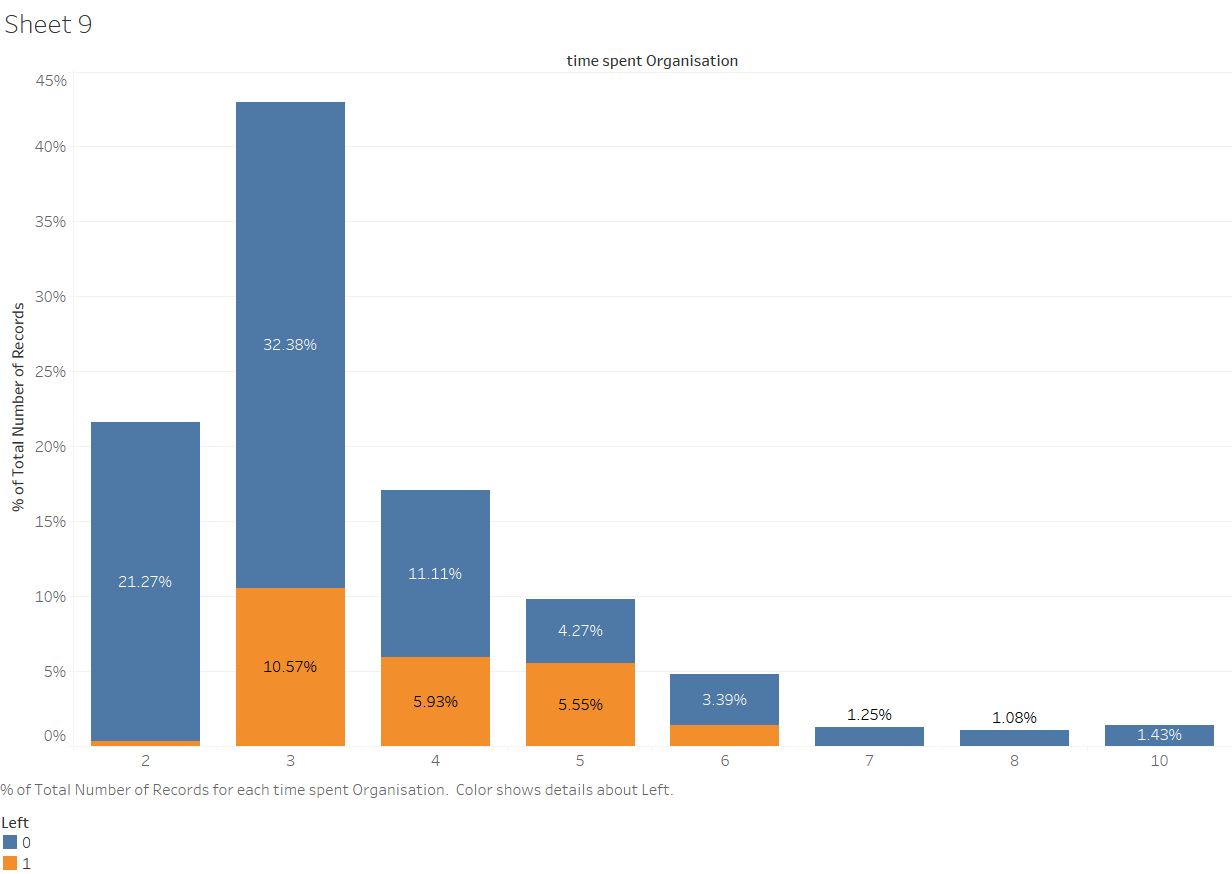
**RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND LAST CR)**

*(***Interpretation***- Personnel leaving are from various CR range. Mostly from .45-.57 and .77-1. 167 out of 283 have left, who have been given 100% CR )*



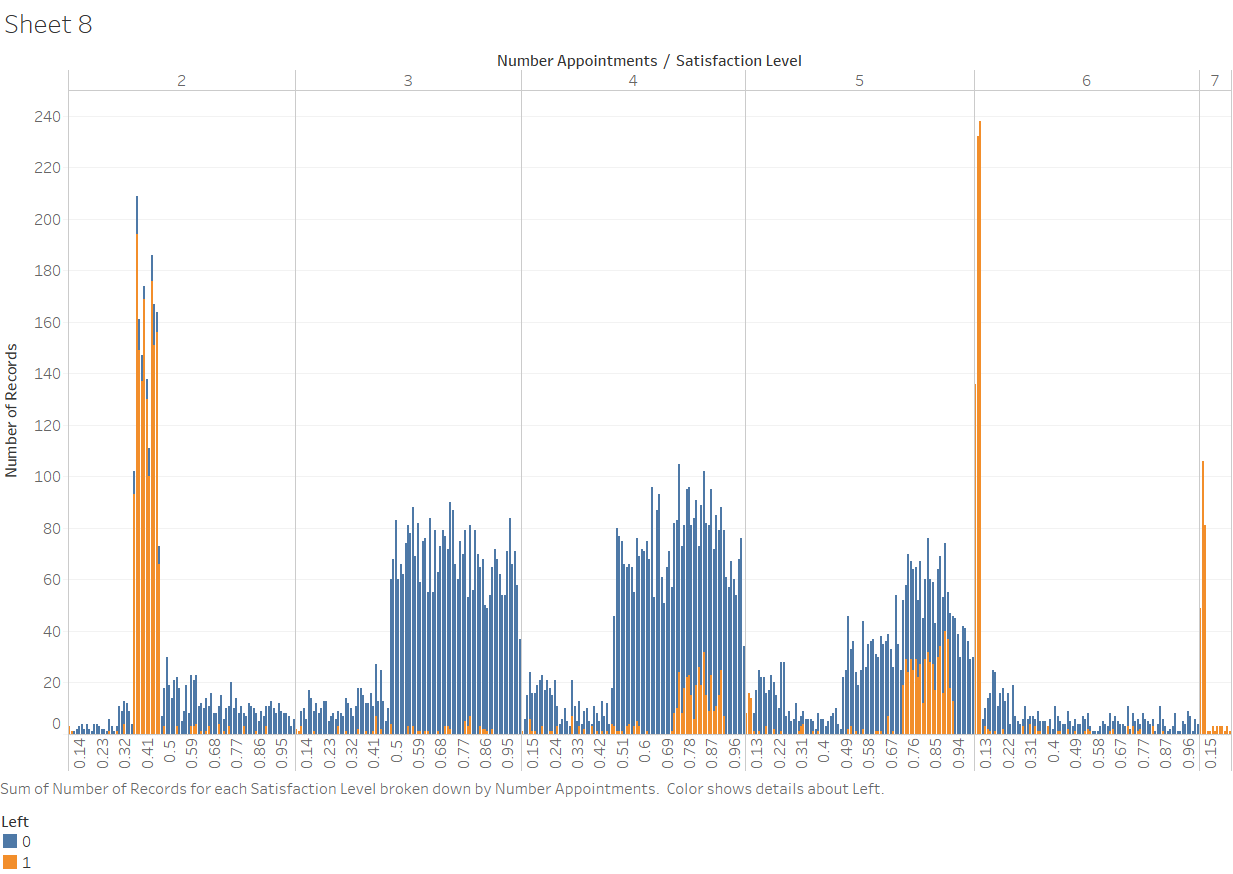
**RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND SATISFACTION LEVEL)**

*(***Interpretation***- Most of the unsatisfied personnel have left. Also, there are large number of personnel left who have satisfaction level between .36-.46)*



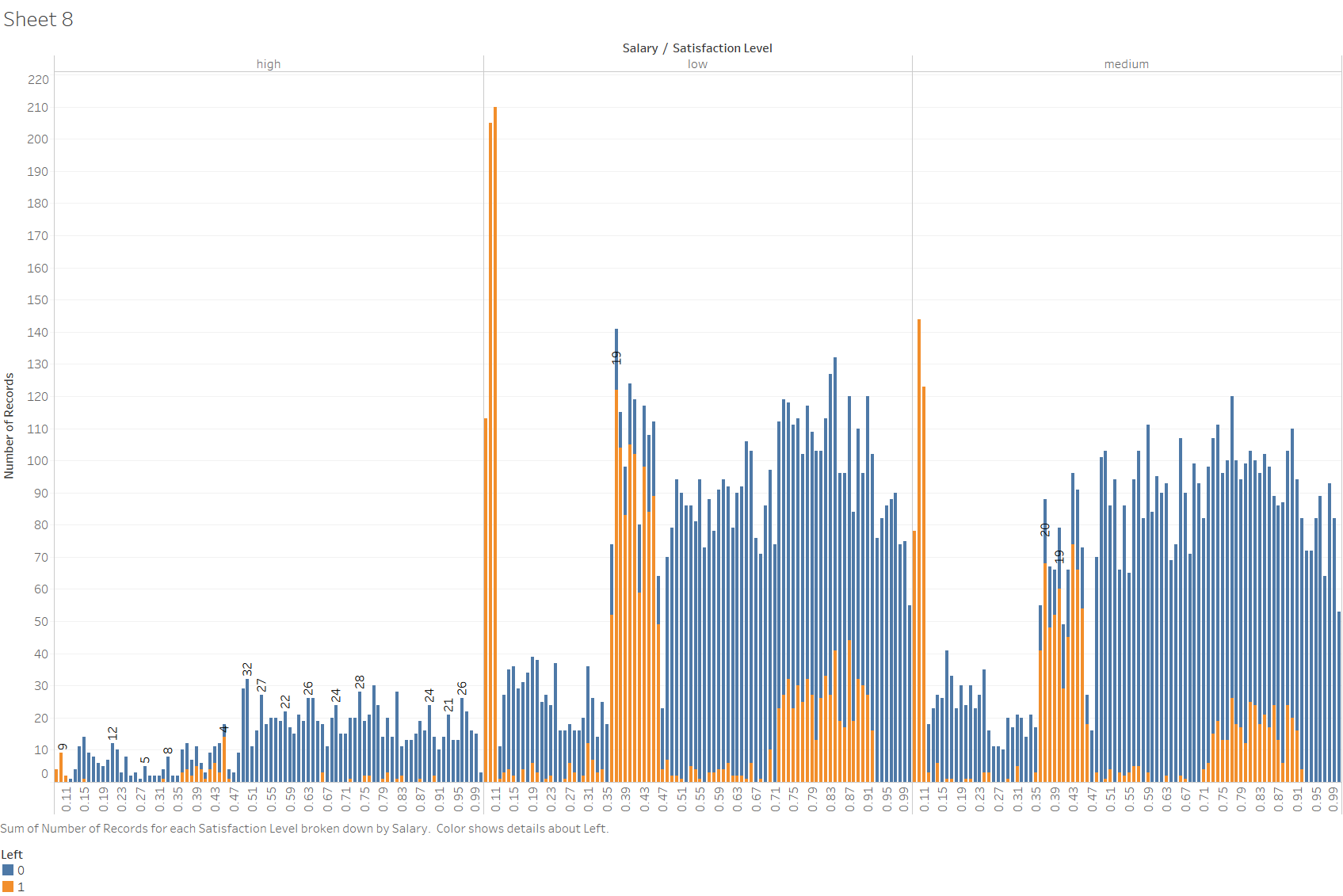
**RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND TIME SPENT IN THE ORGANISATION)**

*(***Interpretation***- There are less number of personnel leaving within 2Yrs of service. Also, personnel who have stayed in the organisation beyond 6 yrs, hardly leave.* *There are 6,123 employees having spent more than 3 years within the company and evaluations higher than 0.7 and 30.44% (1,864 employees) of them have left the company)*



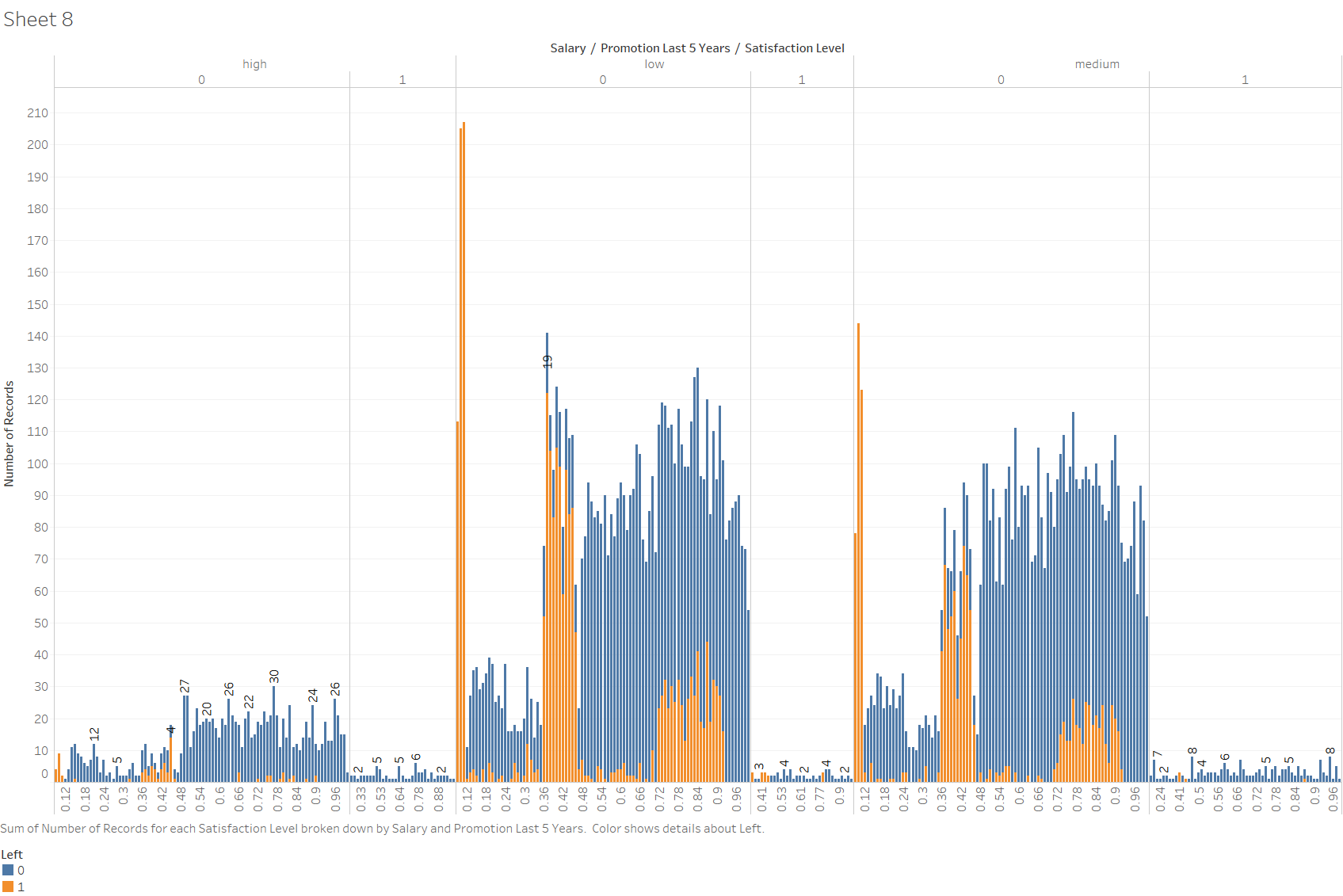
**RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND NO OF SIMULTANEOUS APPOINTMENTS(PROFILES) / PROJECTS AND SATISFACTION(0-1)**

*(***Interpretation***- Personnel who have less or very large(>6) no of simultaneous projects, seems to have less satisfaction and have left the organisation. The other clusters of orange in the graph may have been due to some policy decision of the firm and may be considered to be the outliers)*



**RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT), SALARY(3- HIGH, 2-MEDIUM, 1-LOW) AND SATISFACTION (0-1 )**

*(***Interpretation***- Personnel in low and medium salary group with low satisfaction level have left the organisation. The other clusters of orange in the graph may have been due to some policy decision of the firm)*

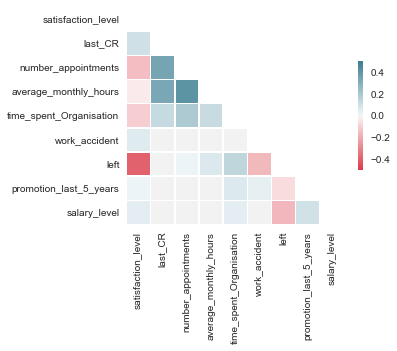


**RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT), SALARY(3- HIGH, 2-MEDIUM, 1-LOW) AND PROMOTION IN LAST 5 YRS**

*(***Interpretation***- Personnel in low and medium salary group with low/medium satisfaction level have left the organisation, if not promoted in last five years.)*

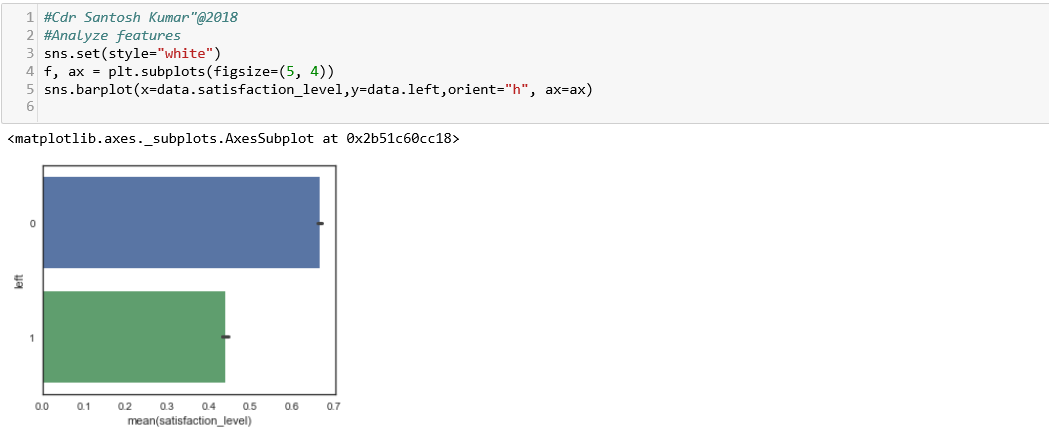
**Data Preparation**

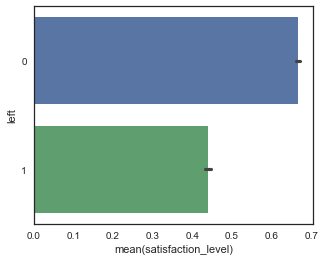
1. **Analyse correlations.** Calculating the correlations between all different combinations of data allows us to get first hints on why people leave in order to orient our analysis into the right perspective. Red fields mean negative correlations, blue fields indicate positive correlations ie. The field on the crossing point between “left” and “satisfaction\_level” is dark red which means that when the satisfaction level of employees goes down, the value of “left” goes up (which means that employees are leaving, as satisfaction\_level can only be 0 or 1).



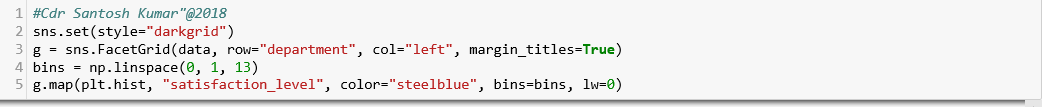


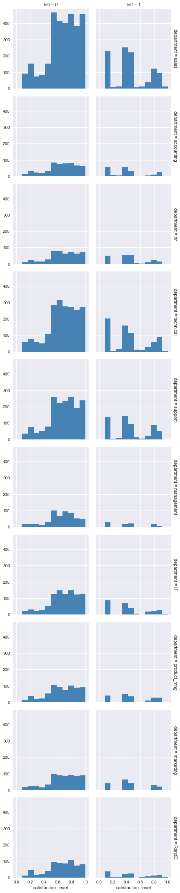
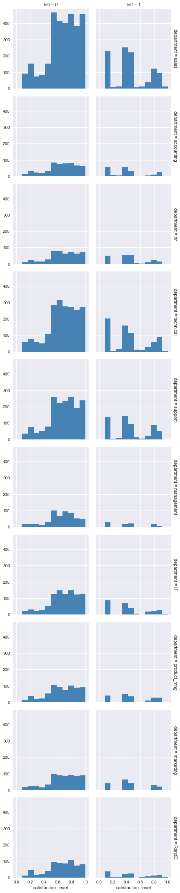
1. From the plot, it can be seen very clearly that the **satisfaction level of the employees is strongly related to the fact that they leave the organisation**. Other significant factors making people leave are the **salary level, the work accidents and if they have got a promotion during the last 5 years**. Regarding the correlation between the satisfaction level and the other dimensions, it can be understood, **that satisfaction mainly decreases when the number of simultaneous appointments(profiles)/ projects and the time spent in the company increase.**
2. **Focus on employee satisfaction.** The mean satisfaction level of personnel leaving is less(.45) as compared to personnel still in the organisation(.65).





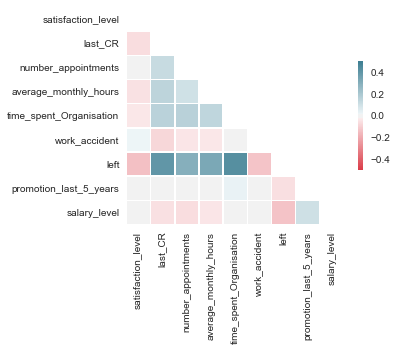
1. We plot histogram for how employee satisfaction looks like for the different departments. In the left range is the charts for employees still in the company (left=0), in the right range is the the employees that have already left the company (left=1).



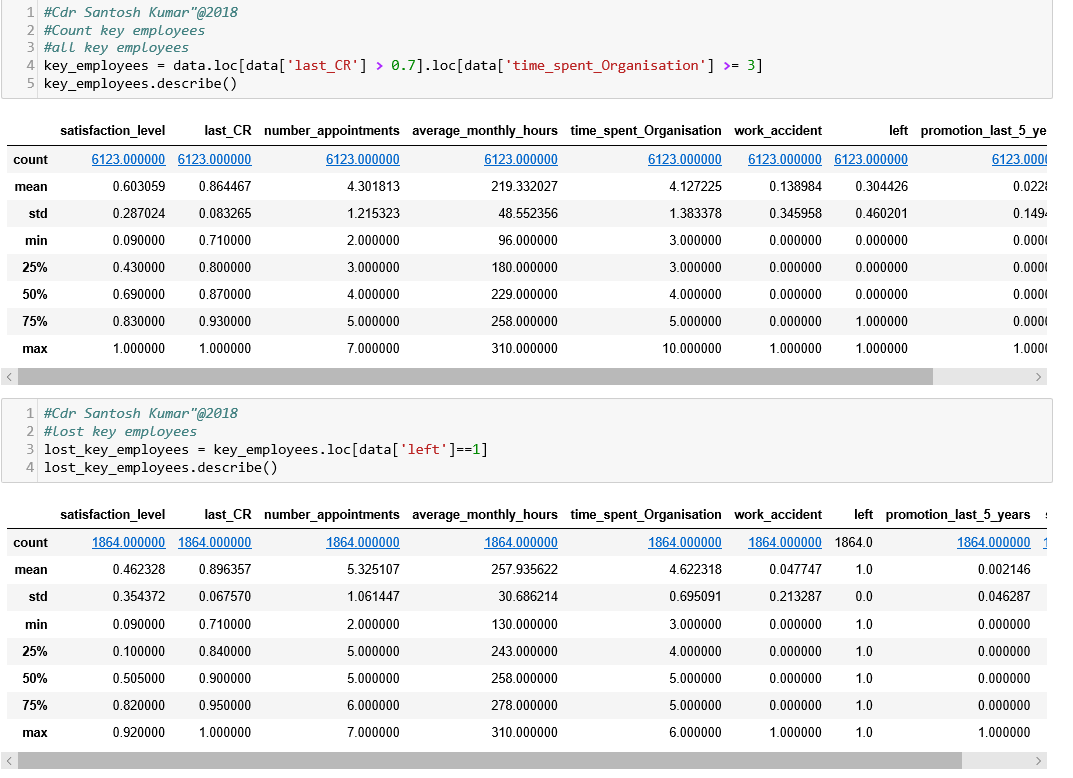


1. Employees that have left can be split up into 3 distinct groups; those who were unsatisfied, those who were very satisfied and those in between. There is no smooth transition between those groups like there is for employees still in the company. It appears quite clearly why unsatisfied people leave the company, but it could be interesting to explore why satisfied employees left. Therefore, we plot the correlation chart including satisfied employees only (satisfaction\_level > 0.7).



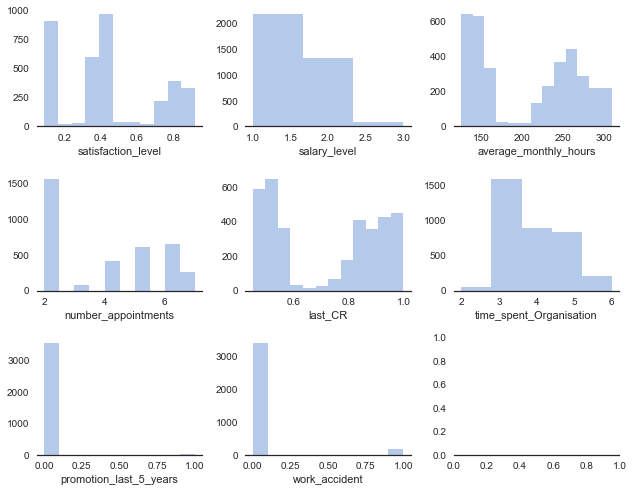


1. This chart shows that **satisfied employees leave the company when they work on a high number of appointments/ projects or a high number of hours each month and when they have already spent a long time in the company**. Leave decisions are also influenced by a low salary level and when employees haven’t got a promotion during the last 5 years.

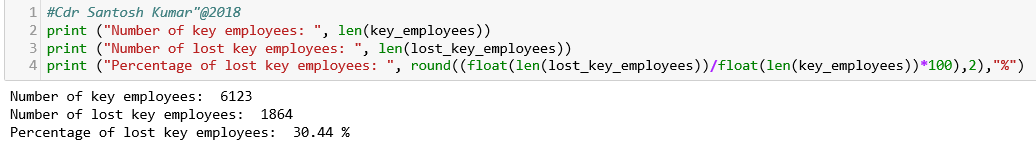
****

1. **Other factors**. We plot histogram for various parameters to analyse  on other factors that describe leaving employees.



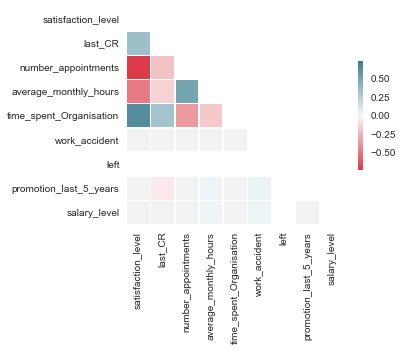


1. It can be seen that leaving employees tend to have lower salaries, a higher number of projects, higher monthly working hours and fewer promotions. All this sounds logic as the satisfaction analysis provides the same conclusions and satisfaction is closely related to the leave decision. It is to be noted that a **large number of the employees leaving the company are people with a high evaluation and several years spent in the company**. These employees are highly valuable assets that should not be lost.There are 6,123 employees having spent more than 3 years within the company and evaluations higher than 0.7 and 30.44% (1,864 employees) of them have left the company.

****

1. **Why do good employees leaving?** Before trying to predict which people are most likely to leave the organisation, it is important to understand what makes high performers leave.





1. This correlation matrixshows that good employees have left mainly because of a high number or simultaneous projects and a high amount of working hours.

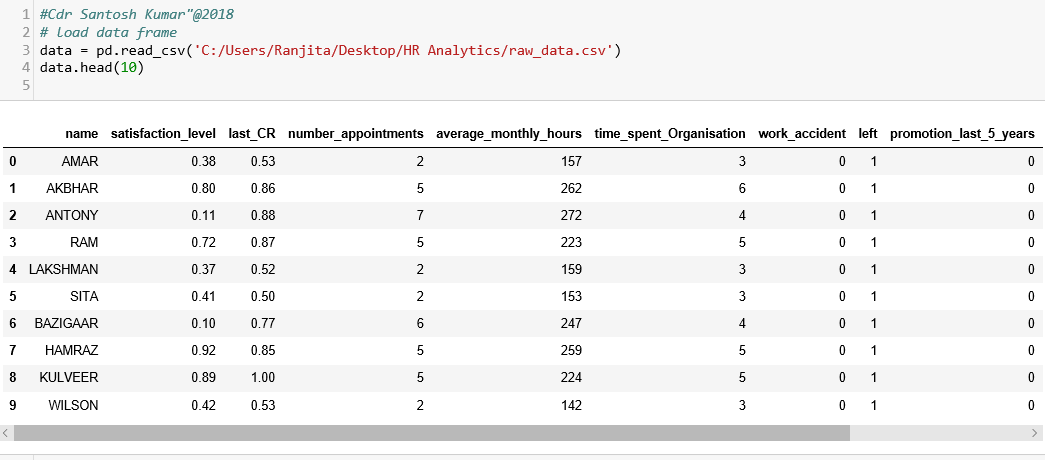
**Section V - Use Case**

**Using Machine Learning Models Predict Which Employees would Leave the Company**

1. In the first part of our analysis, some plots have been generated to get some basic insights about data set and the features have showed quite good correlation rates. In this part data prepared would be used to predict which employees will leave the company. A model needs to be created to predict extremely accurately which employees will leave the company and who will stay. The precision level of the model has to be high.



1. Loading all 15000 lines of the data into a data frame and display the column names as well as the first five rows of the data set.

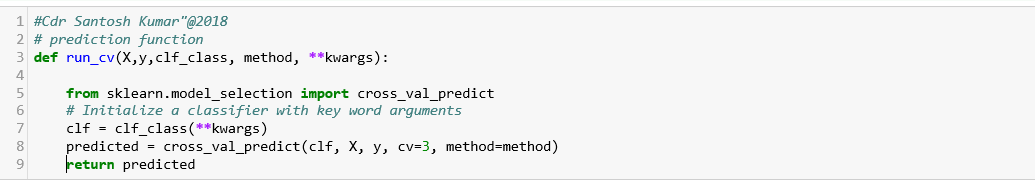


1. **Feature Preparation**. Feature preparation is required for feature standardization. After having transformed the department labels into integers and put all values into a feature matrix, feature standardization converts all values into floats ranging between -1.0 and 1.0. This procedure contributes considerably to the accuracy of the predictions.

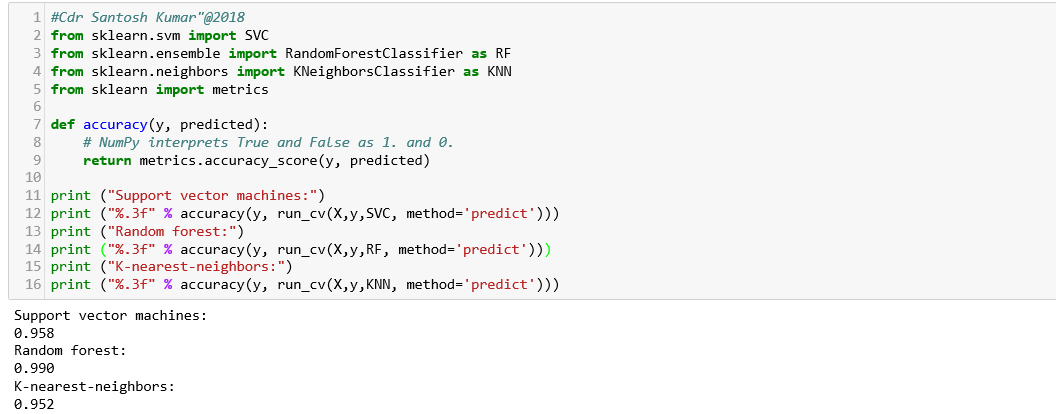


**Data Modelling**

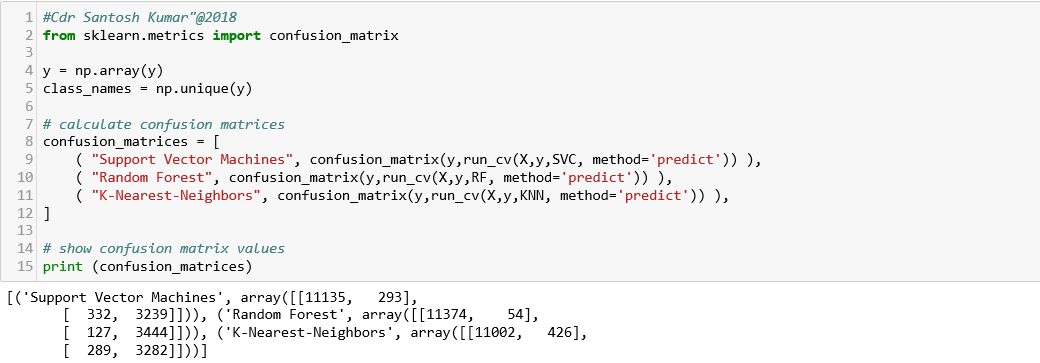
1. **Prediction function.** The prediction function uses a 3-fold cross-validation and it takes several prediction algorithms as input (we would use it to compare 3 different algorithms). We call this function both for comparing the algorithms and predicting the leave probabilities.



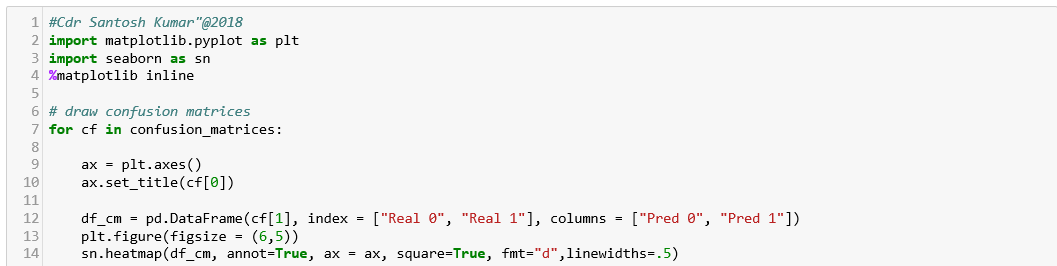
1. **Prediction Algorithms.** The following classification techniques would be evaluated:-
   1. **Support Vector Machine (SVM).** It is a supervised machine learning algorithm which can be used for both classification or regression challenges. However,  it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.[[21]](#footnote-21)
   2. **Random Forest Classifier.** It creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.[[22]](#footnote-22)
   3. ***K*-Nearest Neighbors Algorithm** (***k*-NN**). It is a  non- parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. [[23]](#footnote-23)
2. **Compare prediction algorithms.** In order to chose the best classification algorithm we need to compared a Support Vector Classifier (SVC), a Random Forest Classifier (RF) and a K-Nearest-Neighbors classifier (KNN). All models produce quite satisfying results (between 95% and 99% accuracy) without even tuning the algorithms. We would retain the Random Forest Classifier for the prediction as it produces the best results (98.8% accuracy).

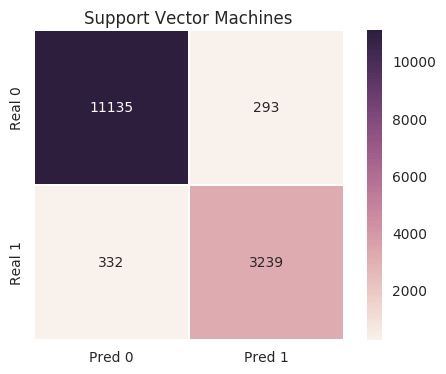


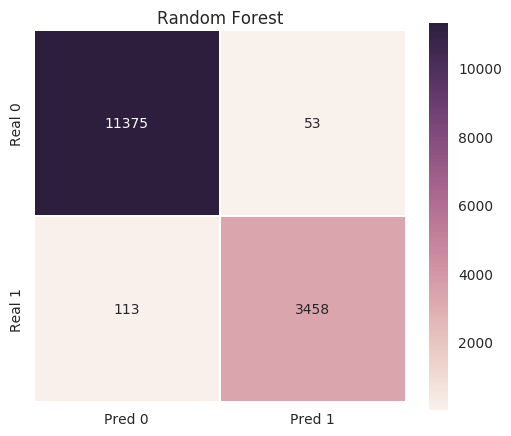
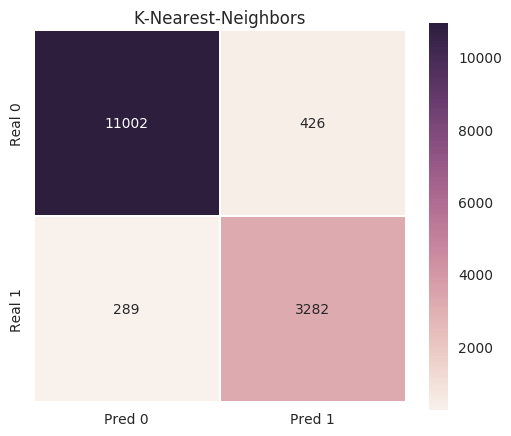
1. **Calculate confusion matrices.** A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known**.** Confusion matrices are a very useful tool to get an overview of the accuracy of a prediction. The matrix provides a value for each crossing point between predicted and realized classes. The confusion matrix has 4 fields: Left-Left, Left-Not left, Not left-Left and Not left-Not left. The Left-Left and Not left-Not left fields contain by far the largest amount of values. It indicates high quality of the prediction.



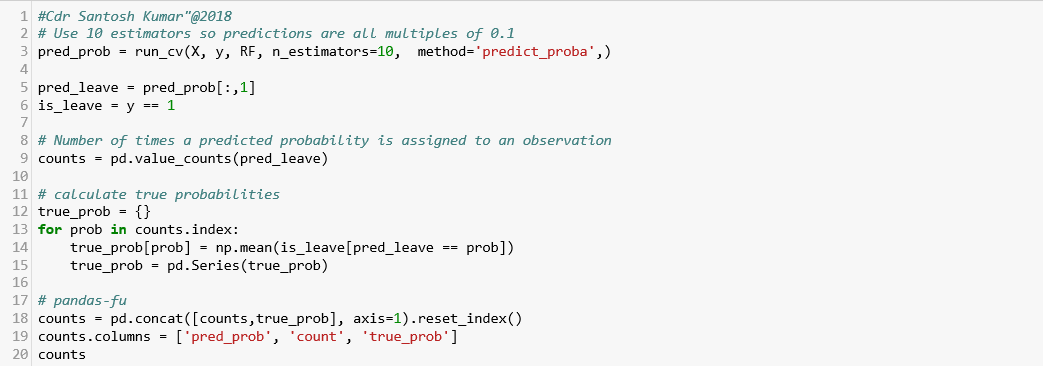
1. **Plotting confusion matrices.** Using Seaborn visualization library of Python we plot the confusion matrices for 3 prediction algorithms. The Random Forest Classifier model predict 11,375 times correctly that an employee will stay.

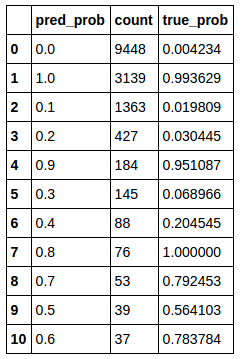


[](http://www.verteego.com/wp-content/uploads/2016/12/svm.png)

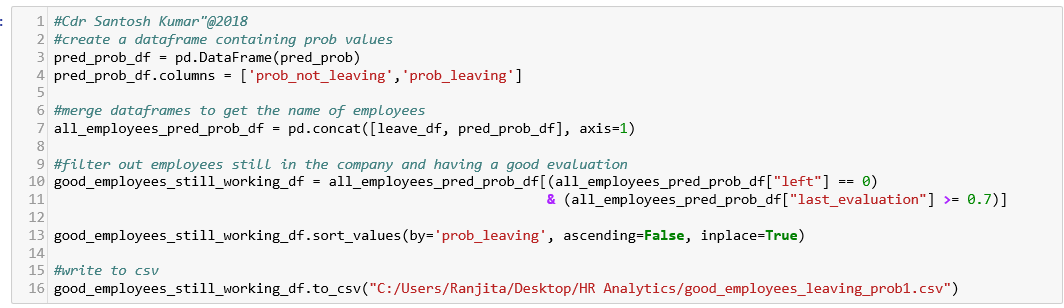
[](http://www.verteego.com/wp-content/uploads/2016/12/rf.png)[](http://www.verteego.com/wp-content/uploads/2016/12/knn.png)

1. **Calculating Prediction Probabilities for all Employees.** We use the predict function to calculate the probabilities for staying and leaving (left=0, left=1) for all 15000 employees in our data set. As every predictor makes several hundreds or even thousands of predictions we can compare our probabilities with the actual outcome of each class.For example, for the group of employees for which we predicted a 60% probability of leaving, the actual leaving percentage is 78%. As we can see the predicted probabilities for the two main classes (pred\_prob=0% and pred\_prob=100%) are very close to the real probabilities which shows another time that our model is extremely accurate.



[](http://www.verteego.com/wp-content/uploads/2016/12/Screenshot-from-2016-12-06-22-19-12.png)

1. **Generate key employees with leaving/ staying probabilities.** In the last step we filter out all key employees (employees that have a last evaluation higher than 0.7) that are still in the company. This gives us a table of about 6000 employees. In order to be able to alert managers about employees that are most likely to leave we order the employee list by their leaving probability and save the whole list as a CSV file.

****

**Applicability of Data Analytics in Naval Civilians**

.

1. In Indian Navy, the sanctioned strength of Naval Civilian is approx. 46000, which constitute about 35% of the total manpower of the Navy, and are poised to grow substantially in the next 15 Yrs. This makes imperative to have an effective and robust Human Resource Management System. The civilian workforce forms an integral part of the service, and their health, well-being, satisfaction, and morale has a direct impact on the combat/ operational readiness of the Navy.
2. **Requirement of Data Ware House**. Using Data Analytics and Predictive Machine Learning, a large number of issues can be resolved and preemptive actions can be initiated to prevent occurrences in the future. However, this can be achieved only if we have substantial database to carry out predictions with assumed hypothesis. To undertake such exercise, there is a requirement of having a Data Warehouse, which becomes a master repository of database of all Naval Civilians.
3. At present, the only reason, for not undertaking such analytics and predictions is the lack of authenticity/ correctness of such database existing with various units. Naval Civilian Management Information System (NCMIS) is a progressive step towards achieving such objective of creating a centralised database of all Naval Civilians.
4. Human Capital would always remain the most vital and potent resource of any organization. An effective Human Capital can only be attained by persistent efforts to address the challenges being faced by an organisation and consistently strive to achieve goals set out by higher leadership. To understand the challenges, we need to explore Data Analytics for early predictions of such issues.

1. <https://www.techopedia.com/definition/28334/human-resources-analytics-hr-analytics> [↑](#footnote-ref-1)
2. <https://www.expertsystem.com/machine-learning-definition/> [↑](#footnote-ref-2)
3. <https://www.techemergence.com/machine-learning-in-human-resources/> [↑](#footnote-ref-3)
4. <https://www.peoplematters.in/article/techhr16/scope-of-robotics-and-artificial-intelligence-in-hr-13806> [↑](#footnote-ref-4)
5. *Shearer C., The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13—22.* [↑](#footnote-ref-5)
6. *Undertaking by the author* [↑](#footnote-ref-6)
7. *Fitz-enz & Mattox, 2014* [↑](#footnote-ref-7)
8. *Pfeffer & Sutton, 2006; Schwarz & Murphy, 2008* [↑](#footnote-ref-8)
9. *Becker, 1964* [↑](#footnote-ref-9)
10. *Fitz-enz & Mattox, 2014* [↑](#footnote-ref-10)
11. *Fitz-enz & Mattox, 2014* [↑](#footnote-ref-11)
12. *Angrave et al. (2016* [↑](#footnote-ref-12)
13. *Fitz-enz & Mattox, 2014* [↑](#footnote-ref-13)
14. *Fitz-enz & Mattox, 2014* [↑](#footnote-ref-14)
15. *Angrave et al., 2016* [↑](#footnote-ref-15)
16. *Davenport, 2006* [↑](#footnote-ref-16)
17. *Schwarz and Murphy (2008),* [↑](#footnote-ref-17)
18. *Fitz-enz and Mattox (2014),* [↑](#footnote-ref-18)
19. *Angrave et al., 2016* [↑](#footnote-ref-19)
20. https://en.wikipedia.org/wiki/Cross-industry\_standard\_process\_for\_data\_mining [↑](#footnote-ref-20)
21. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ [↑](#footnote-ref-21)
22. https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1 [↑](#footnote-ref-22)
23. https://en.wikipedia.org/wiki/K-nearest\_neighbors\_algorithm [↑](#footnote-ref-23)