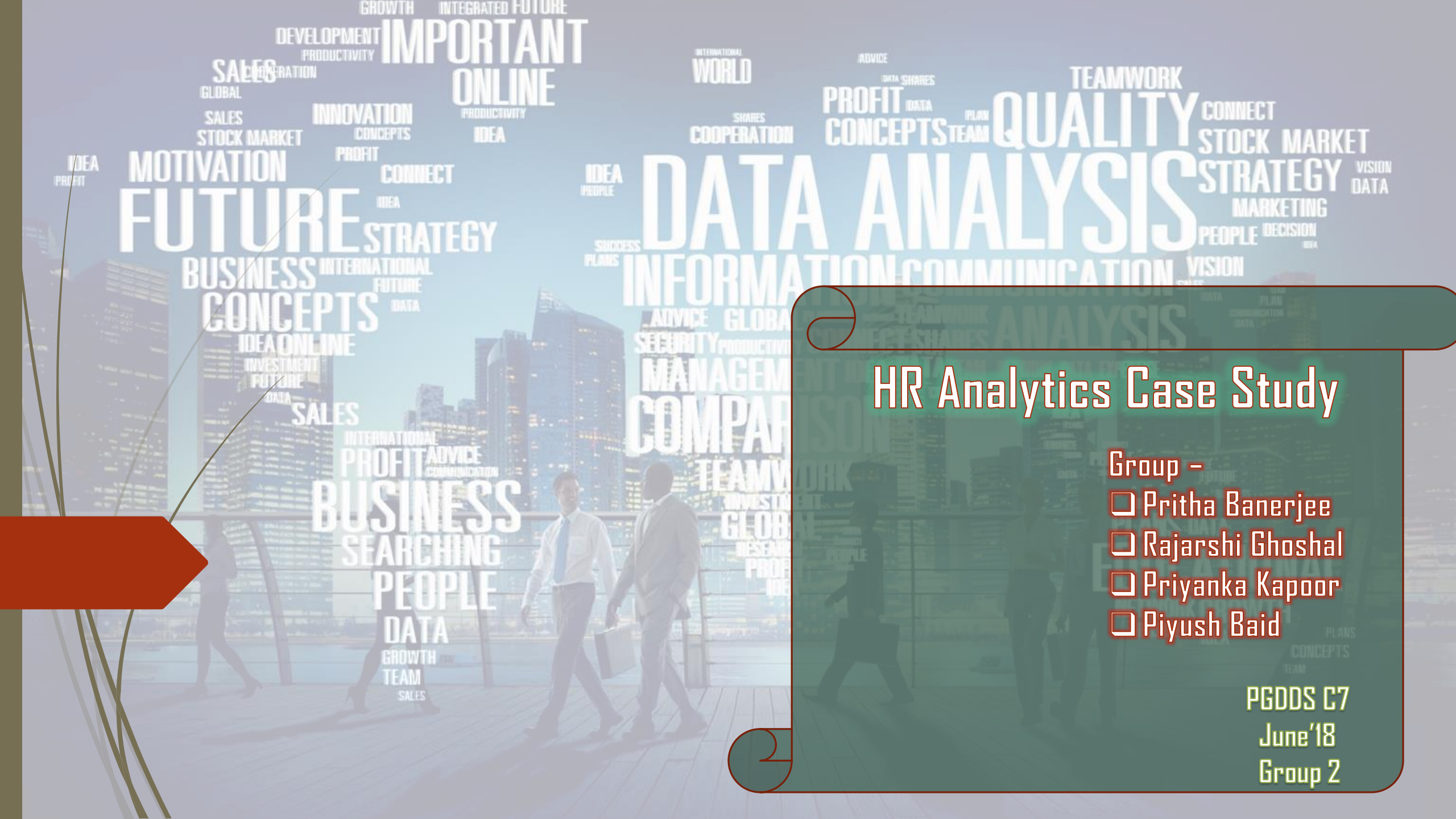


HR Analytics Case Study

Group –

- Pritha Banerjee
- Rajarshi Ghoshal
- Priyanka Kapoor
- Piyush Baid

PGDDS C7
June'18
Group 2

Business Understanding

A large company named **XYZ** has contracted the **HR analytics** firm to understand what factors they should focus on, in order to curb attrition. The company's **Attrition Rate** (employees leaving, either on their own or because they got fired) **is 15% per year** and those employees needs to be replaced with the talent pool available in the job market.

- **High Attrition results in –**
  - The former employees' projects get delayed.
  - Difficult to meet **Project timelines.**
  - Reputation loss among consumers and partners.
  - Maintenance cost behind a large dept. fro recruiting new talent.
  - Further additional training and development expenses.
  - New employees needs to be given a grace period for acclimatizing themselves to the company

Goals and Methodology

The Company **XYZ** wants to know what changes they should make to their workplace, in order to get most of their employees to stay.

- **Our goals, as an HR Analytics firm, are as follows –**
  - Model the probability of Attrition using a Logistic Regression.
  - Determine the variables effecting the Attrition Rate.
  - Pin down the changes to be made to their workplace for the Employee Retention.
  - Suggest how to curb the Attrition rate base on the findings

Metadata and Data Understanding

- There are **4410 unique observations** with respect to the employees collected from the companies survey database
- **Employee Survey Data –** Data based on the survey of Employee's satisfaction over different parameters (3).
- **Manager Survey Data –** Data based on the survey of Manager's feedback over different parameters (2).
- **In_time and Out_time Data –** Data based on Employee's **In** and **Out time** over a period of 12 months.
- **General Data (24 variables) –** Various Employee Information, such as ID, Age, Department, Education, Job Role, Marital Status, etc.
- No Duplicate data nor any missing values were observed with the data set provided.

Assumptions

- Columns in **In_time and Out_time Data** with all **NA values** are considered to be **Mandatory Holidays**, and are thereby removed, such as – "2015.01.01" / "2015.01.14" / "2015.01.26" / "2015.03.05" / "2015.05.01" / "2015.07.17" / "2015.09.17" / "2015.10.02" / "2015.11.09" / "2015.11.10" / "2015.11.11" / "2015.12.25"
- **Average working hours in a day** per employee over a period of 12 months is assumed to be the best figure for modelling and predictions.
- **NA values** of few of the columns are replaced with the **Median values**, such as – **"EnvironmentSatisfaction", "JobSatisfaction", "WorkLifeBalance", "NumCompaniesWorked"** and **"TotalWorkingYears**".
- Removing 3 variables from data frame which have the same value for all rows – "EmployeeCount", "Over18" and "StandardHours".

Problem Solving Methodology -

Business Understanding

- Getting a clear picture of the business requirement

Extracting the Data

- Downloading / Importing the Data into R

Data Preparation

- Cleaning and formatting the Data into the desired form

Data Selection

- Understanding the Data visually and then, Filter and screen the relevant data

Deriving Metrics

- Variables derived as an Inference and used for analyzation

Exploratory Data Analysis

- Performing Univariate and Bivariate Analysis

Model Building

- We use Predictive Analysis to build Model based on **VIF (variance inflation factor)** and **P value (significance)**

Evaluation

- The Model is evaluated based on **its** discriminative power - **Accuracy, Sensitivity, Specificity, Gain, Lift and KS Statistics**

Review

- Review the Analysis and provide Final Recommendation

Data Preparation and EDA

- All the data sets were cleaned and prepared to be merged to form a core data file for analysis.
- NA values for few of the numerical variables were replaced with median values.
- Derived metrics, such as **Average Working hours, Number of leaves taken** were created.
- Dummy variables were created for the categorical variables having 2 or more factor levels.:

  2 Levels – Gender, Attrition, Performance Ratings

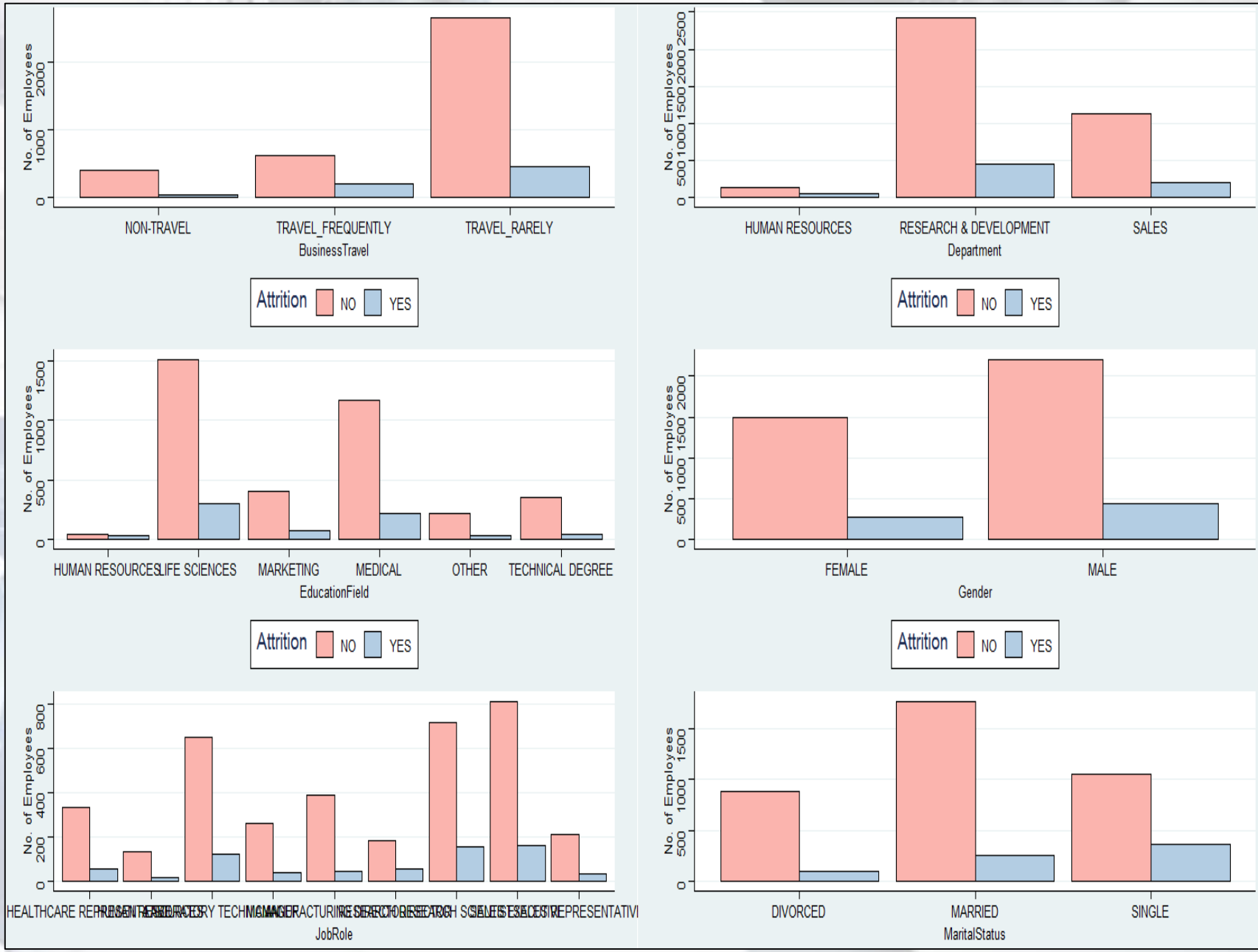
  More than 2 Levels – Environment Satisfaction, Job Satisfaction, Work Life Balance, Job Involvement, Business Travel, Department, Education, Education Field, Job Level, Job Role, Marital Status, Stock Option Level.
- Few of the **variable's scales were standardised** so that it doesn't have a disproportionate effect on the model's results.
- Outliers were treated as well for **Monthly Income, Total Working Years, Years At Company, Years Since Last Promotion and Years With Curr Manager.**
- Columns having constant values were removed – **Employee Count, Over 18** and **Standard Hours;**

Derived Metrics

- **Average Work Hours** – Average of the Daily Working hours over a year's data was derived out of the 2 data sets – **In_time** and **Out_time.**
- **Number of leaves** – Apart from Weekends and Mandatory holidays, if an employee has NA values for the In_time date column, he/she is assumed to be on leave for that particular day. Sum of all those NA values for each employee is derived as their number of leaves in a year.

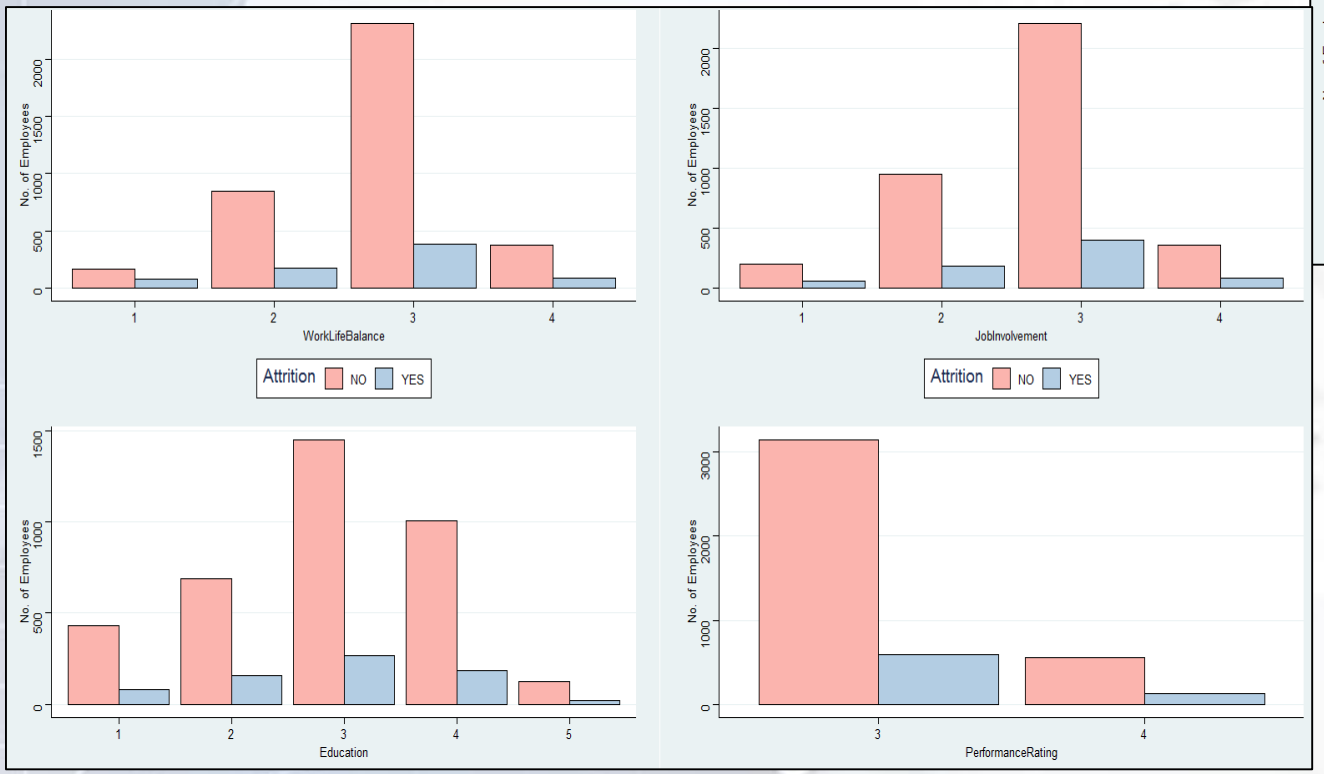
Visualization – Part 1

- **Business Travel**– Those who travel rarely have a higher attrition.
- **Department**– People working in research and development department have a higher tendency of attrition.
- **Education Field**– People from Life sciences field have a higher attrition rate.
- **Gender**– Male show higher attrition compared to females.
- **Job Role**– Research Scientist and sales executive show higher attrition compared to others
- **Marital Status**– Single have higher rate than others



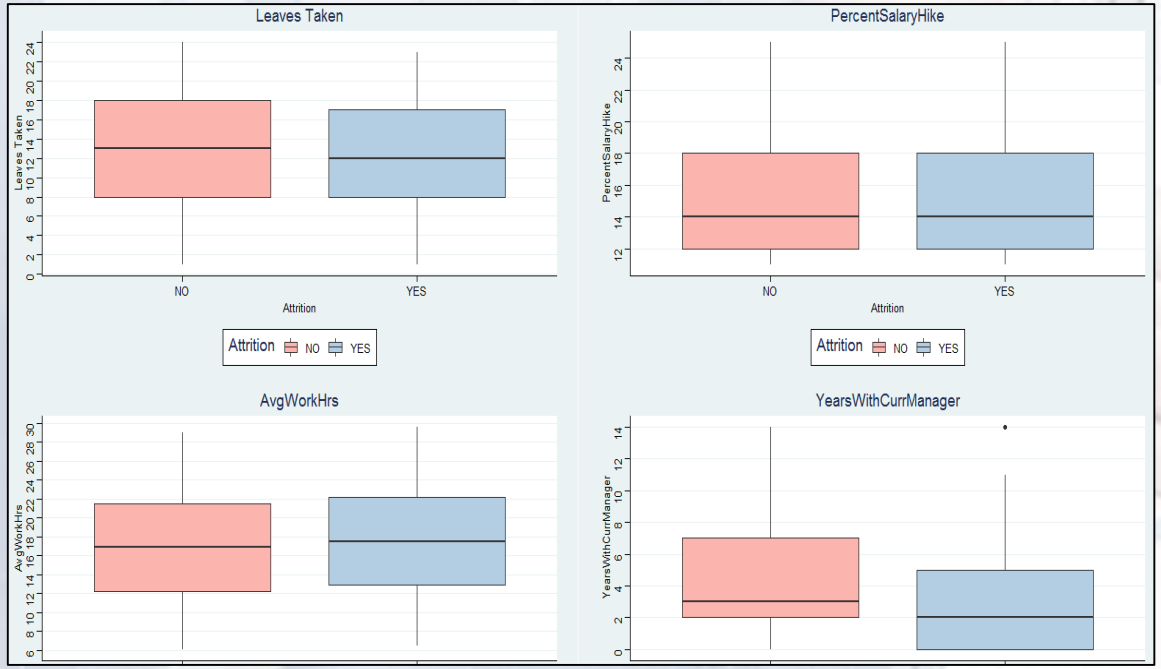
Visualization-Part 2

- **Environment Satisfaction**– Attrition is high when value is 1.
- **Job Satisfaction**– People giving rating of 3 show higher tendency of attrition.
- **Work Life Balance–** surprisingly people rating “Better”(3) show higher chances of attrition,

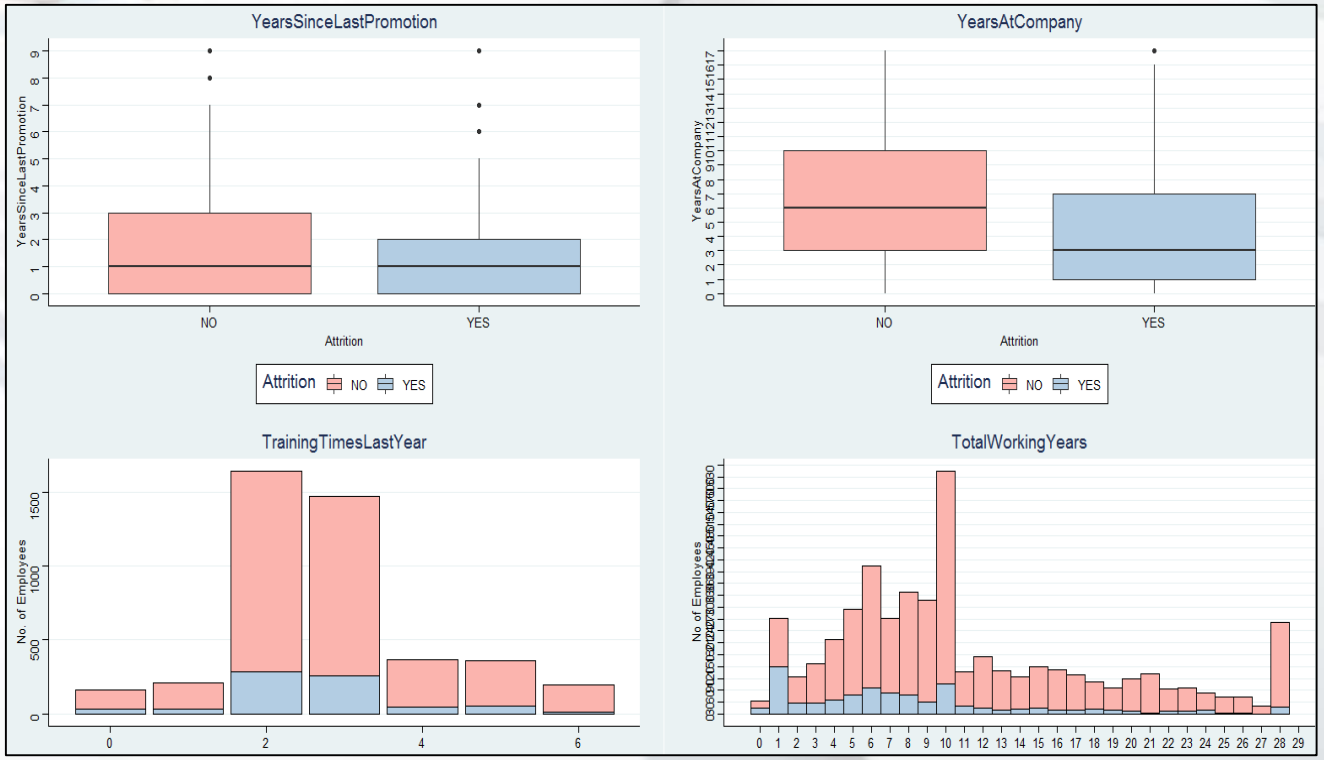


- **Job Involvement**– People who are highly involved in their jobs have higher attrition rate.
- **Education**– People who done “Bachelors” tend to attrition more.
- **Performance Rating–** People given a rating of 3 show higher chances of attrition.

Visualization – Part 3



- **Leaves Taken**– It shows a similar pattern, nothing much can be inferred.
- **Percent Salary Hike** It shows a similar pattern, nothing much can be inferred
- **AvgWorkHrs–** Those who work more hours on an average show higher attrition.
- **Years With Curr Manager-** working for a less time with current manager show higher tendency of attrition.



- **Years Since Last Promotion**– Those who have not been promoted in the last 2 years will attrition more.
- **Years At Company/Total Working Years**– People who are associated with less years with the company show higher tendency of attrition.
- **Training Times Last Year–** People who have been trained for 2/3 times last year show higher chances of attrition.

Model Building and Evaluation

- Once the data is cleaned and prepared, it is divided into 2 parts for Model building-

  Training data is used for the model to learn during modelling.

  Testing data is used by the trained model for prediction and model evaluation.

- **GLM function** was used to get the Initial Model and then with the help of **Step AIC**, we received a standard model to begin our modelling with.

- Based on **VIF (variance inflation factor)** and **P value (with significance),** we reduced the variables and achieved our final model with almost all predictors being significant with lowest VIF are present.

- We then deduced the measures of discriminative power of a model- **Accuracy, Sensitivity, Specificity, Gain, Lift and KS Statistics**. A high KS statistic means that not only does your model have all churns at the top, it has all non-churns at the bottom. For a good model, **KS statistic would be more than 40% and would lie in the top few deciles (1st to 4th).**

Logistic Regression Model

- With the help of the Logistic Regression, final working model was prepared to predict the attrition rate for the company.

```
Final Model ← glm(formula = Attrition ~ Age + NumCompaniesWorked +  TotalWorkingYears +
                  TrainingTimesLastYear + YearsSinceLastPromotion +  YearsWithCurrManager +
                  EnvironmentSatisfaction2 + EnvironmentSatisfaction3  + EnvironmentSatisfaction4 +
                  JobSatisfaction4 + WorkLifeBalance2 +  WorkLifeBalance3 + WorkLifeBalance4 +
                  BusinessTravelTRAVEL_FREQUENTLY + BusinessTravelTRAVEL_RARELY +
                  JobRoleMANUFACTURING.DIRECTOR + MaritalStatusSINGLE ,
                  family = "binomial", data = train)
```

A total of 17 variables were found to be significant as per the model which are relevant for the prediction

Model Evaluation- Discriminative Powers of a Model

With the help of Confusion Matrix, we evaluated the Sensitivity, Specificity and Accuracy for different Cut-off Probability.

Optimal Cut-Off equalizes accuracy, sensitivity and specificity.

Cut-Off Probability	Sensitivity	Specificity	Accuracy
50%	0.173913	0.9775986	0.8518519
40%	0.3043478	0.9444444	0.8442933
30%	0.4541063	0.890681	0.8223734
Otimal Cut-Off (16.95%)	0.7004831	0.703405	0.7029478

Model Assessment- Gain, Lift and KSS

Decile	Observations	Churn	Cum-Churn	% Cum-Churn(Gain)	Cum-Lift	Non-Churn	Cum-Non-Churn	%Cum-Non-Churn	KSS (%Cum-Churn) - (%Cum-Non-Churn)
1	133	66	66	31.9%	3.19	67	67	6.0%	25.9%
2	132	43	109	52.7%	2.63	89	156	14.0%	38.7%
3	132	31	140	67.6%	2.25	101	257	23.0%	44.6%
4	133	15	155	74.9%	1.87	118	375	33.6%	41.3%
5	132	18	173	83.6%	1.67	114	489	43.8%	39.8%
6	132	12	185	89.4%	1.49	120	609	54.6%	34.8%
7	133	9	194	93.7%	1.34	124	733	65.7%	28.0%
8	132	5	199	96.1%	1.20	127	860	77.1%	19.1%
9	132	7	206	99.5%	1.11	125	985	88.3%	11.3%
10	132	1	207	100.0%	1.00	131	1116	100.0%	0.0%
Total	1323	207				1116			

The KS statistic shows that the model is good in distinguishing between employees who will leave the company and employees who won't because it satisfies both the criteria:

- **is equal to 40% or more**
- **lies in the top deciles, i.e. 1st, 2nd, 3rd or 4th**

Significant Variables and their Coefficients

Coefficients	Estimate	Std. Error	Z Value	Pr (>|Z|)	Significant Codes
Age	-0.30629	0.07609	-4.025	5.69E-05	***
NumCompaniesWorked	0.33954	0.05537	6.132	8.68E-10	***
TotalWorkingYears	-0.45091	0.09862	-4.572	4.82E-06	***
TrainingTimesLastYear	-0.18912	0.05554	-3.405	0.000661	***
YearsSinceLastPromotion	0.48576	0.07054	6.886	5.74E-12	***
YearsWithCurrManager	-0.48235	0.08216	-5.871	4.33E-09	***
EnvironmentSatisfaction2	-0.54168	0.15945	-3.397	0.000681	***
EnvironmentSatisfaction3	-0.66885	0.14665	-4.561	5.10E-06	***
EnvironmentSatisfaction4	-1.01984	0.15129	-6.741	1.57E-11	***
JobSatisfaction4	-0.7343	0.12663	-5.799	6.68E-09	***
WorkLifeBalance2	-0.80932	0.2178	-3.716	0.000202	***
WorkLifeBalance3	-1.08913	0.2029	-5.368	7.97E-08	***
WorkLifeBalance4	-0.95997	0.25195	-3.81	0.000139	***
BusinessTravelTRAVEL_FREQUENTLY	1.84128	0.2757	6.679	2.41E-11	***
BusinessTravelTRAVEL_RARELY	1.09889	0.26208	4.193	2.75E-05	***
JobRoleMANUFACTURING.DIRECTOR	-0.81342	0.21359	-3.808	0.00014	***
MaritalStatusSINGLE	0.87369	0.10949	7.98	1.47E-15	***

Recommendations

- According the variables as per our Final Logistic Regression Model, we arrive at the following observations –

- Better the **Job Satisfaction**, **Work-Life Balance** and **Environment Satisfaction** of the Employee, less are the chances of their attrition.
- If an Employee has worked in multiple companies in the past, the chances of that employee leaving becomes high.
- **Total Working years** of the Employee effects the attrition negatively. More the experience of the candidate, less likely the chances of the employee for attrition.
- If an Employee works with the same manager for a longer time, less chances of him/her to leave the company.
- Employees, who are **Single**, have high probability of leaving the Organization.
- **Promotion** seems to be a positive factor for retaining employees.
- **Tenured Employees** are more likely to stay in the Organization.