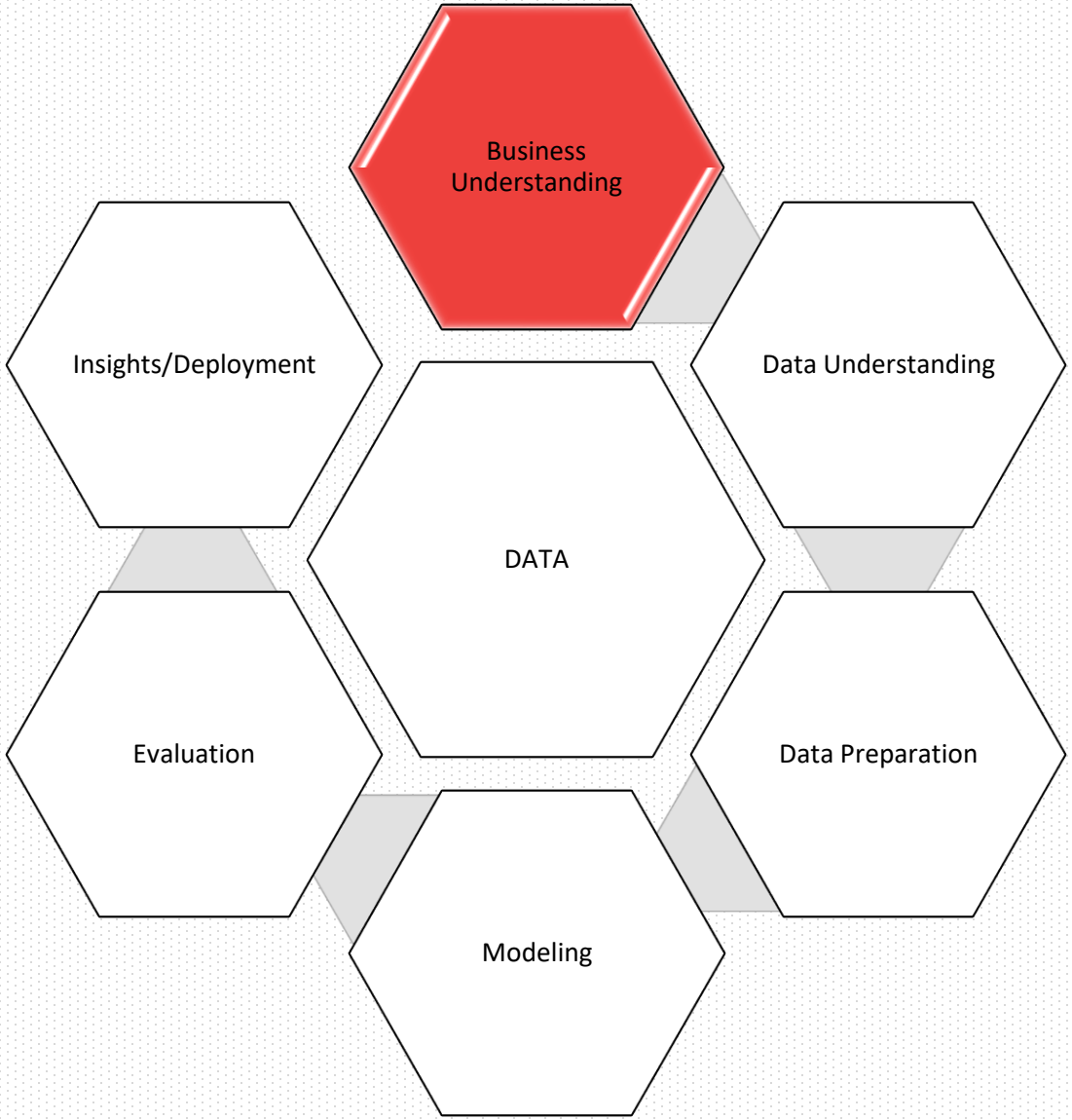
Analysis of Employee Attrition

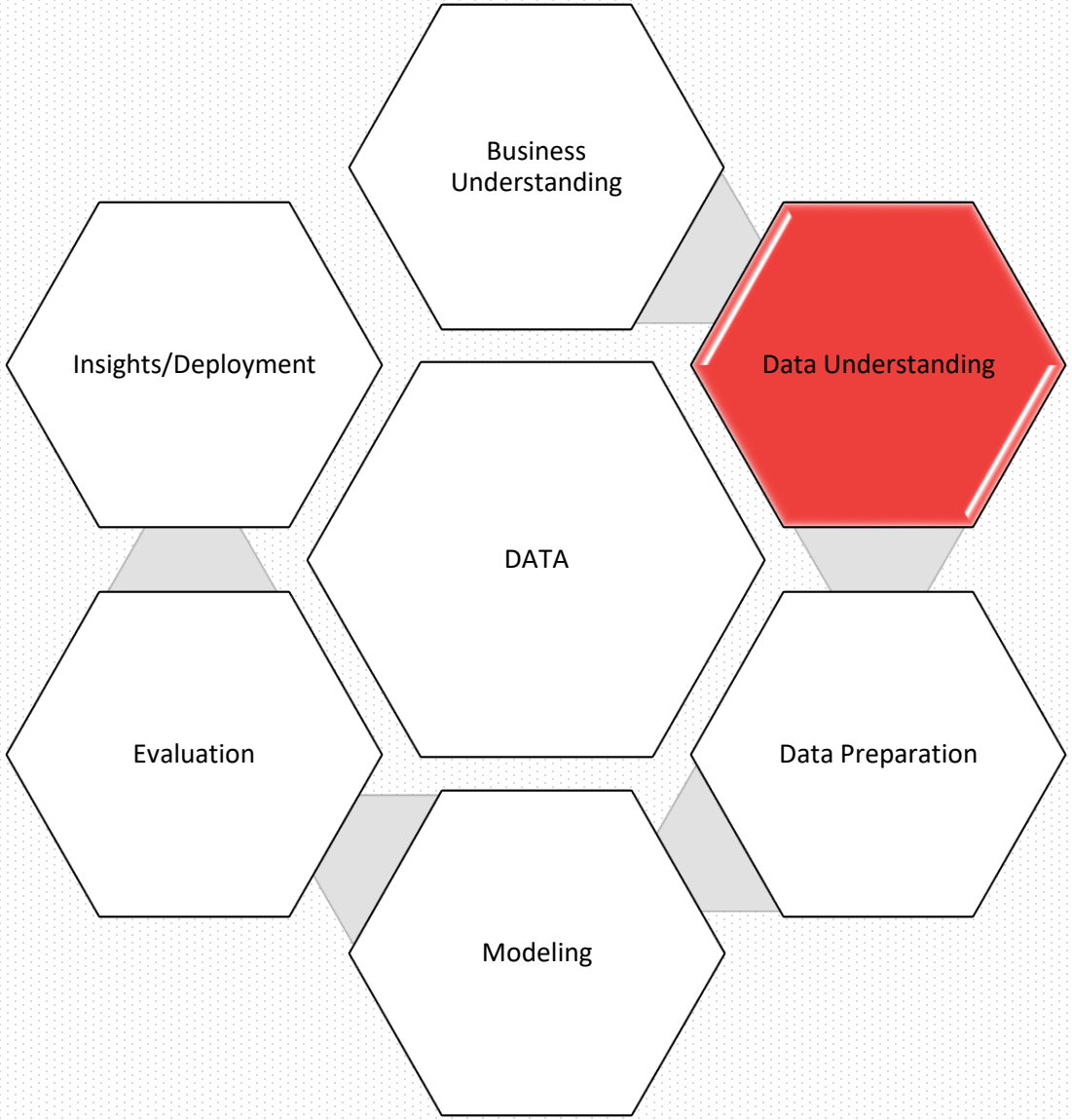
Charwak Apte
Alfred Mburugu
Mridula Kamath
Abdi Adam
03/2018

UpGrad

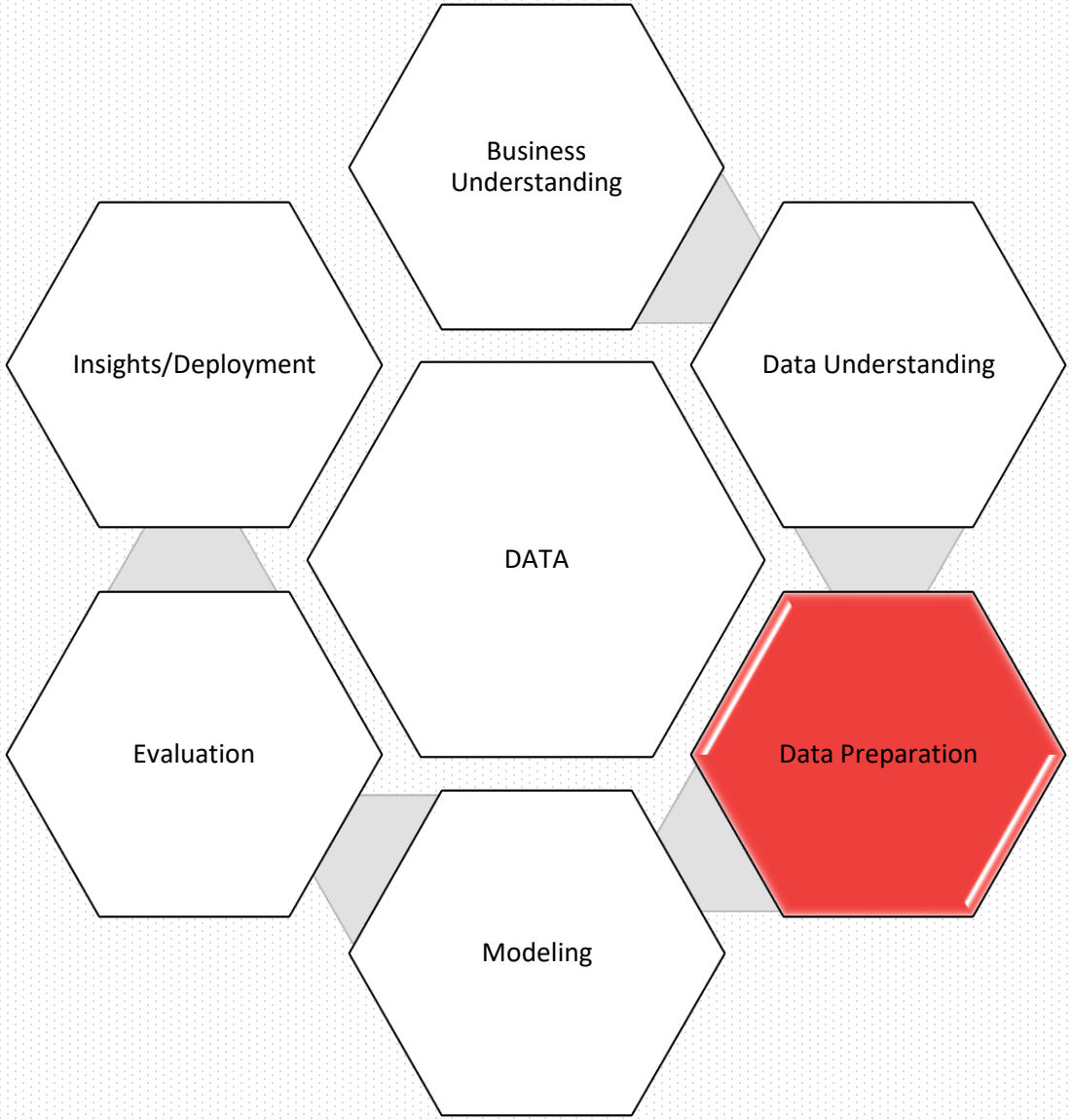
imt·b


Business Understanding
Insights/Deployment
Data Understanding
DATA
Evaluation
Data Preparation
Modeling

- Problem Definition:
  - XYZ faces 15% attrition annually
  - Attrition results in:
    - Delays and inability to hit timelines
    - Investment in hiring personnel.
    - New hires need time to ramp up.
  - Figure out:
    - Factors which influence employee attrition.
    - What changes to be made in the workplace.
    - What needs to be addressed right away.
- Approach:
  - Formulate the attrition probability as a logistic regression.
  - Identify critical factors and action items to inhibit attrition.


Business Understanding
Insights/Deployment
Data Understanding
DATA
Evaluation
Data Preparation
Modeling

- Data Available for 4410 employees, ~30 variables spread across 5 files
  - General Employee Data
    - EmployeeID(Unique Identifier), Age, Education, JobLevel etc (4410 observations w/ 24 variables)
  - Employee Survey Data
    - Environment/Job Satisfaction/Work-Life Balance
  - Manager Survey Data
    - Job Involvement/Performance Rating
  - In time for each employee for a one-year period
    - Provides in-time for each employee on a daily basis for one year.
  - Out-time for each employee for a one-year period
    - Provides out-time for each employee on a daily basis for one year.


Business Understanding
Insights/Deployment
Data Understanding
DATA
Evaluation
Data Preparation
Modeling

- In-Time and Out-Time Handling
  - Convert timestamp strings to date-time variables
  - Use (Out-time – In-time) = WorkTime per day
  - Create a new variable to summarize average worktime of each employee over the year.
- Joined all data frames by EmployeeID as the unique identifier.
- Removed observations with NA values after confirming that data-loss is < 3%.
- Scaled continuous variables to their Z-scores (example: average hours/Percent Salary Hike etc)
- Converted Likter Scale variables to categorical values and created dummy variables. For example, Job Involvement/Performance Rating etc.
- Converted Age from continuous to 3 buckets Early/Mid and Late Career.
- Split the resulting dataset into training and testing data with a 70-30 split.


Business Understanding
Insights/Deployment
Data Understanding
DATA
Evaluation
Data Preparation
Modeling

- Use glm to model with family=binomial to model attrition probability as a function of the initial 57 coefficients.

```
#Initial model
model_1 = glm(Attrition ~ ., data = train[,-1], family = "binomial")
summary(model_1)
#Null deviance: 2661.4  on 3009  degrees of freedom
#Residual deviance: 1990.9  on 2953  degrees of freedom
#AIC: 2104.9
```

- AIC : 2104.9 – This is used as a baseline to compare future simplified models.
- Residual Deviance reduces to 1990.9 from 2661.4 which indicates that the model is certainly doing better than a simple intercept-model and is fitting the attrition phenomenon better.
- However 56 variables add to model complexity and so in the next few slides we will demo simpler models using stepAIC and p-value/vif filtering.

- Use stepAIC to iterate quickly to simplify model_1

```
model_2<- stepAIC(model_1, direction="both")
summary(model_2)
#Null deviance: 2661.4  on 3009  degrees of freedom
#Residual deviance: 2004.8  on 2972  degrees of freedom
#AIC: 2080.8
```

- AIC : 2080.8 (2104 baseline) – A decrease in AIC indicates that the iterative process was able to get a a good fit even with 38 (baseline : 57)
- Residual Deviance reduces to 2004.8 from 2661.4 which indicates that the model is certainly doing better than a simple intercept-model and is fitting the attrition phenomenon better.
- This model has 38 coefficients and is simpler than model_1 but we will simplify it further for a minor loss of accuracy in the model fit (AIC).

- Use p-value check to eliminate statistically non-significant (p-value > 0.05) and also used VIF to get rid of variables with high multicollinearity if they are not statistically significant.

```
#Eliminating these variables with p-values > 0.05
model_3 <- glm(formula = Attrition ~ NumCompaniesWorked + TotalWorkingYears +
                 TrainingTimesLastYear + YearsSinceLastPromotion + YearsWithCurrManager +
                 AvgHrs + Age.xMid + EnvironmentSatisfaction.xLow +
                 EnvironmentSatisfaction.xVHigh + JobSatisfaction.xLow + JobSatisfaction.xVHigh +
                 WorkLifeBalance.xBest + WorkLifeBalance.xBetter + WorkLifeBalance.xGood +
                 BusinessTravel.xTravel_Frequently + BusinessTravel.xTravel_Rarely + EducationField.xMedical +
                 EducationField.xOther + EducationField.xTechnical.Degree + JobLevel.xLevel5 + JobRole.xManager +
                 JobRole.xManufacturing.Director + JobRole.xResearch.Director +
                 MaritalStatus.xMarried + MaritalStatus.xSingle +
                 StockOptionLevel.xStockLevel1, family = "binomial", data = train[,-1])
summary(model_3)
#Null deviance: 2661.4  on 3009  degrees of freedom
#Residual deviance: 2046.8  on 2983  degrees of freedom
#AIC: 2100.8, slight AIC increase but acceptable to reduce model complexity.

vif(model_3)
coef_3 <- coef(summary(model_3))
sort(vif(model_3))
#VIF still low, eliminating p values which are > 0.01 to reduce model complexity:
```

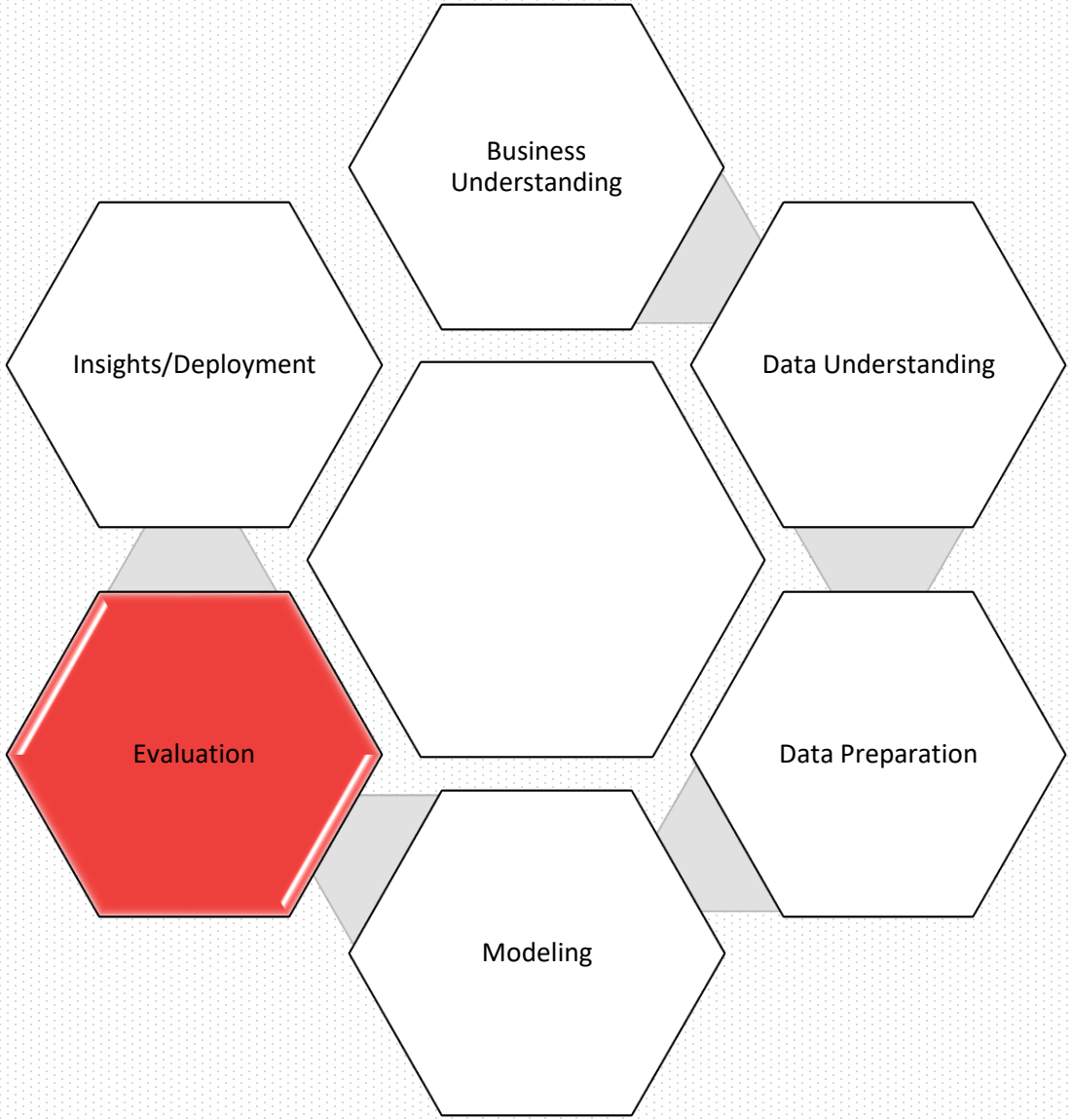
- Since the coefficients are now only 27 (lower from 38 in model_2), listing all of them in the image, it starts giving us a clue about the factors which matter.

- Use p-value check to eliminate more variables to simplify the model. Eliminate variables with p-value > 0.01

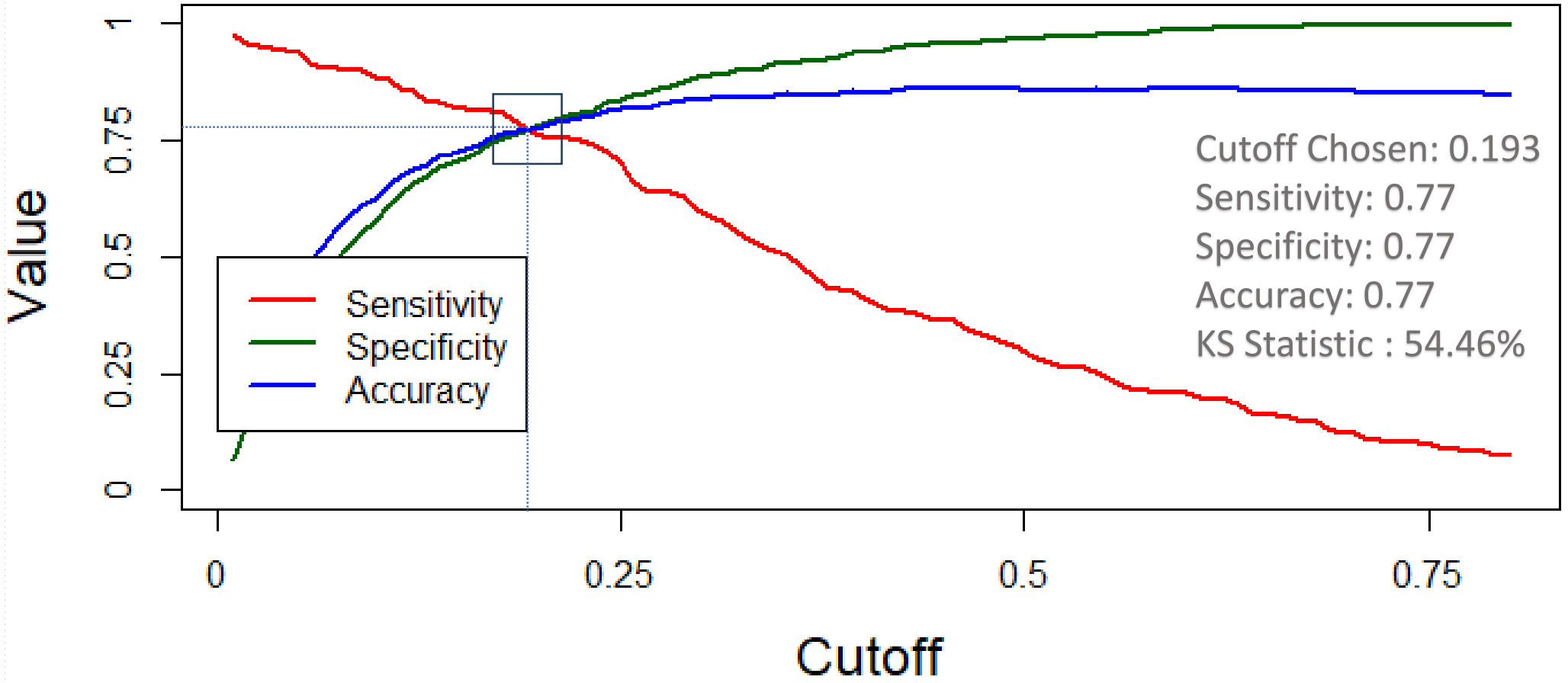
```
model_4 <- glm(formula = Attrition ~ NumCompaniesWorked + TotalWorkingYears +
                 TrainingTimesLastYear + YearsSinceLastPromotion + YearsWithCurrManager +
                 AvgHrs + Age.xMid + EnvironmentSatisfaction.xLow +
                 JobSatisfaction.xLow + JobSatisfaction.xVHigh +
                 WorkLifeBalance.xBest + WorkLifeBalance.xBetter + WorkLifeBalance.xGood +
                 BusinessTravel.xTravel_Frequently + BusinessTravel.xTravel_Rarely +
                 JobRole.xManufacturing.Director + JobRole.xResearch.Director +
                 MaritalStatus.xSingle, family = "binomial", data = train[,-1])

summary(model_4)
#Null deviance: 2661.4  on 3009  degrees of freedom
#Residual deviance: 2077.1  on 2991  degrees of freedom
#AIC: 2115.1
vif(model_4)
coef_4 <- coef(summary(model_4))
sort(vif(model_4))
```

- Since the coefficients are now only 19 (lower from 27 in model_3), listing all of them in the image, it starts giving us a clue about the factors which matter.


Business Understanding
Insights/Deployment
Data Understanding
Evaluation
Data Preparation
Modeling

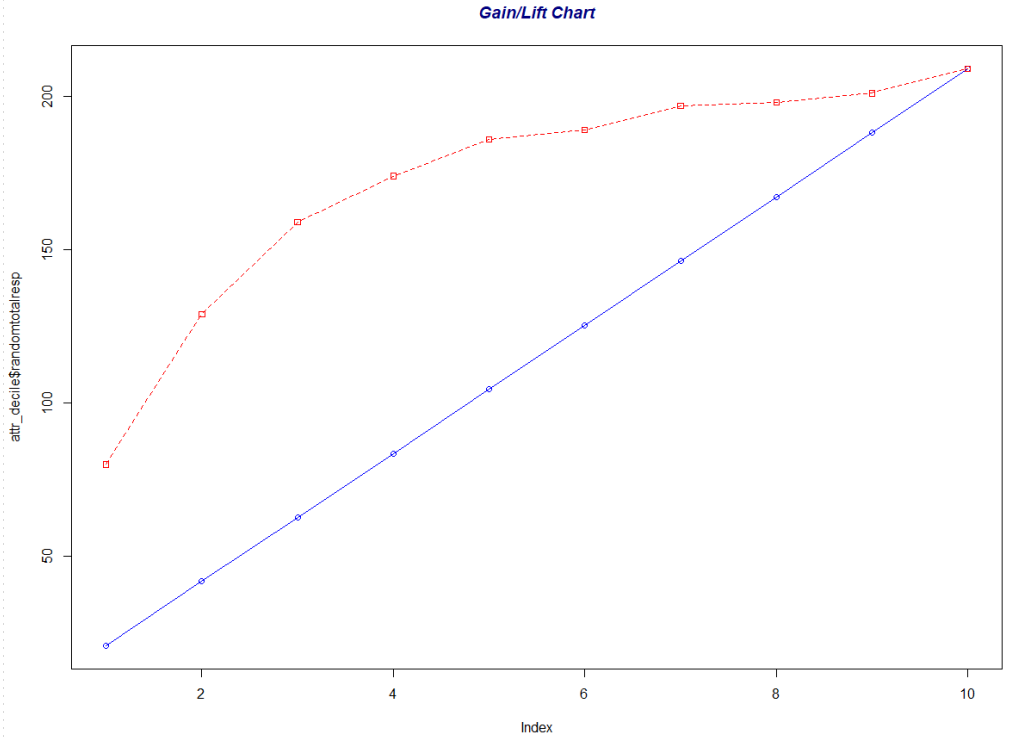
- Look for Probability Cutoff to classify whether or not the employee left.
- We pick the point which maximizes all three attributes.



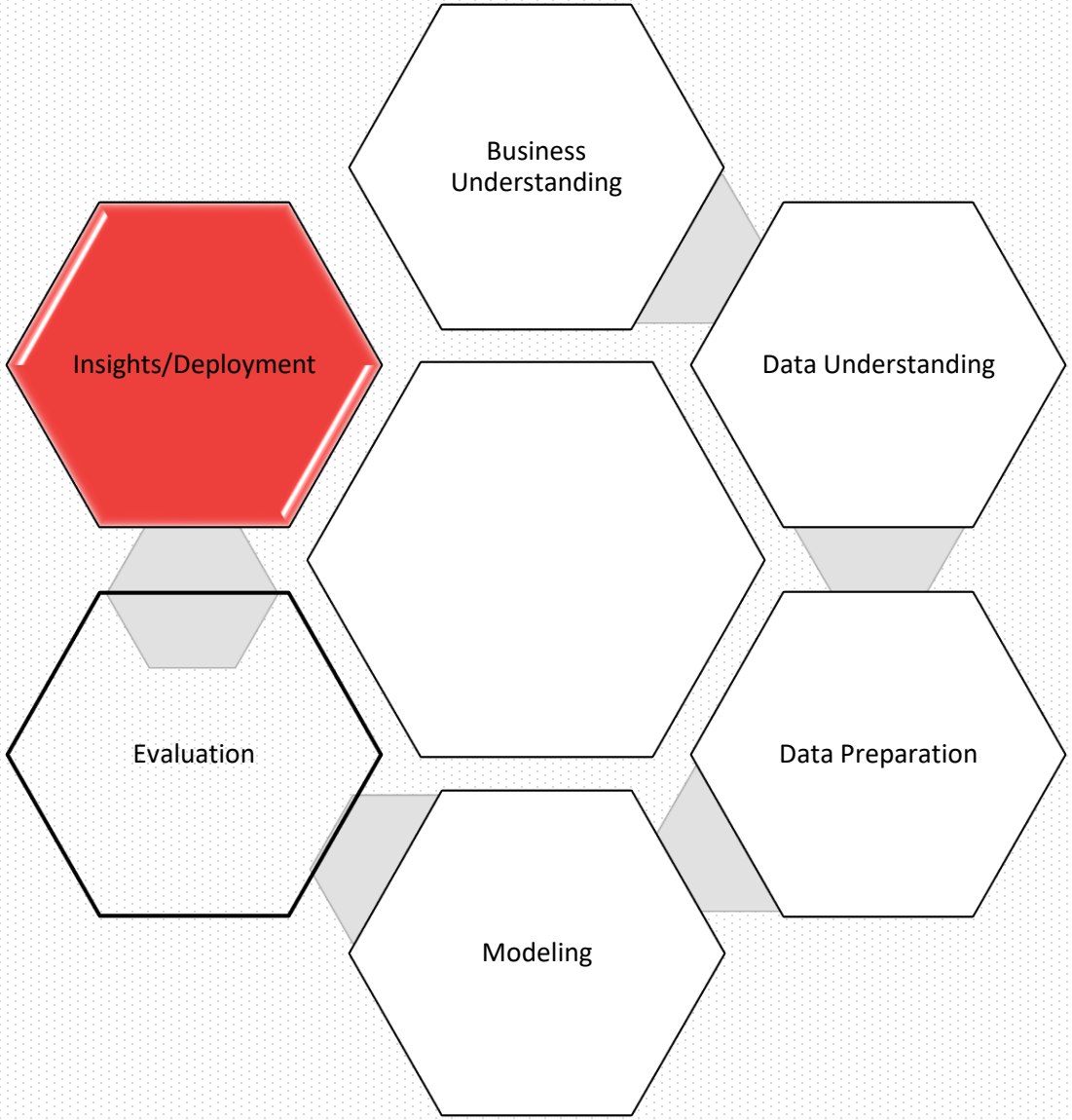
- KS-Statistic is computed to be : 54.56%, occurs in the 5th decile.

- Lift- Gain charts
- The red line indicates the model covers 83.25% of the employees lost in
  - The first 4 deciles of the population.

	bucket	total	totalresp	Cumresp	Gain	Cumlift	randomtotalresp
1	1	129	80	80	38.27751	3.827751	20.9
2	2	129	49	129	61.72249	3.086124	41.8
3	3	129	30	159	76.07656	2.535885	62.7
4	4	129	15	174	83.25359	2.081340	83.6
5	5	129	12	186	88.99522	1.779904	104.5
6	6	129	3	189	90.43062	1.507177	125.4
7	7	129	8	197	94.25837	1.346548	146.3
8	8	129	1	198	94.73684	1.184211	167.2
9	9	129	3	201	96.17225	1.068581	188.1
10	10	129	8	209	100.00000	1.000000	209.0



- Area Under red curve/Area under the blue curve = 1.5


Business Understanding
Insights/Deployment
Data Understanding
Evaluation
Data Preparation
Modeling

- Use model_4 as final model. Since coefficient count is manageable and gives a decent fit.

Coefficients Ordered by Magnitude

Higher values Reduce attrition

Higher values Worsen attrition

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9518853	0.29831178	-6.543105	6.025446e-11
WorkLifeBalance.xBetter	-1.2260279	0.21113795	-5.806762	6.369239e-09
WorkLifeBalance.xBest	-1.0017544	0.26762255	-3.743161	1.817194e-04
WorkLifeBalance.xGood	-1.0015756	0.22634958	-4.424906	9.648432e-06
TotalWorkingYears	-0.7457147	0.09015163	-8.271783	1.319704e-16
JobRole.xManufacturing.Director	-0.7367082	0.21669535	-3.399742	6.744951e-04
Age.xMid	-0.6210897	0.11821921	-5.253712	1.490641e-07
YearsWithCurrManager	-0.5002488	0.08744051	-5.721019	1.058870e-08
JobSatisfaction.xVHigh	-0.4935803	0.13867155	-3.559348	3.717769e-04
TrainingTimesLastYear	-0.2166358	0.05862001	-3.695594	2.193731e-04
NumCompaniesWorked	0.2793989	0.05858663	4.768987	1.851548e-06
JobRole.xResearch.Director	0.5909296	0.22596853	2.615097	8.920218e-03
JobSatisfaction.xLow	0.6495600	0.14014367	4.634958	3.570102e-06
AvgHrs	0.6557353	0.05385894	12.175050	4.221351e-34
YearsSinceLastPromotion	0.6626118	0.07799666	8.495387	1.972748e-17
BusinessTravel.xTravel_Rarely	0.6648641	0.22768557	2.920097	3.499219e-03
EnvironmentSatisfaction.xLow	0.8148328	0.13241653	6.153558	7.576350e-10
MaritalStatus.xSingle	0.9584458	0.11642381	8.232386	1.835203e-16
BusinessTravel.xTravel_Frequently	1.2759289	0.24756154	5.153987	2.550063e-07

Coefficients Ordered by p-values

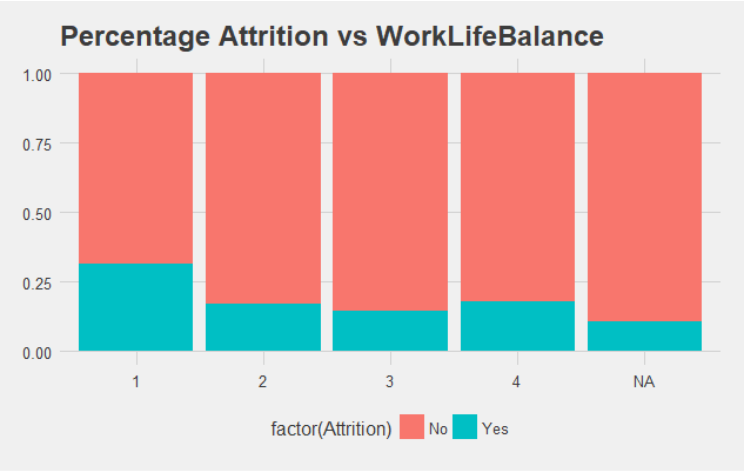
	Estimate	Std. Error	z value	Pr(>|z|)
AvgHrs	0.6557353	0.05385894	12.175050	4.221351e-34
YearsSinceLastPromotion	0.6626118	0.07799666	8.495387	1.972748e-17
TotalWorkingYears	-0.7457147	0.09015163	-8.271783	1.319704e-16
MaritalStatus.xSingle	0.9584458	0.11642381	8.232386	1.835203e-16
(Intercept)	-1.9518853	0.29831178	-6.543105	6.025446e-11
EnvironmentSatisfaction.xLow	0.8148328	0.13241653	6.153558	7.576350e-10
WorkLifeBalance.xBetter	-1.2260279	0.21113795	-5.806762	6.369239e-09
YearsWithCurrManager	-0.5002488	0.08744051	-5.721019	1.058870e-08
Age.xMid	-0.6210897	0.11821921	-5.253712	1.490641e-07
BusinessTravel.xTravel_Frequently	1.2759289	0.24756154	5.153987	2.550063e-07
NumCompaniesWorked	0.2793989	0.05858663	4.768987	1.851548e-06
JobSatisfaction.xLow	0.6495600	0.14014367	4.634958	3.570102e-06
WorkLifeBalance.xGood	-1.0015756	0.22634958	-4.424906	9.648432e-06
WorkLifeBalance.xBest	-1.0017544	0.26762255	-3.743161	1.817194e-04
TrainingTimesLastYear	-0.2166358	0.05862001	-3.695594	2.193731e-04
JobSatisfaction.xVHigh	-0.4935803	0.13867155	-3.559348	3.717769e-04
JobRole.xManufacturing.Director	-0.7367082	0.21669535	-3.399742	6.744951e-04
BusinessTravel.xTravel_Rarely	0.6648641	0.22768557	2.920097	3.499219e-03
JobRole.xResearch.Director	0.5909296	0.22596853	2.615097	8.920218e-03

- Since the range of all variables is controlled by variable scaling, the magnitude of the variables closely indicates sensitivity of attrition to a specific variable.
- Positive Valued coefficients indicate factors that affect attrition adversely (higher log odds and hence probability of attrition), negative coefficients indicate that these factors help retain employees better.

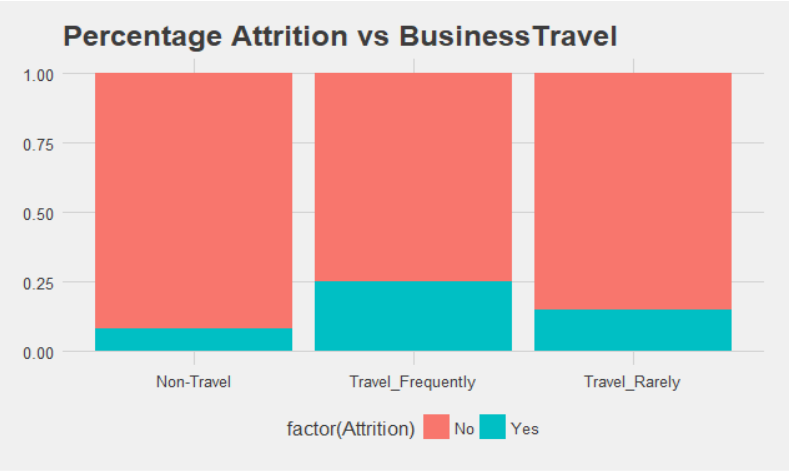
FACTORS INFLUENCING ATTRITION BASED ON THE LOGISTIC REGRESSION MODEL:

- Business Travel (employees who travel frequently have a higher attrition probability)
- Work-Life Balance (employees with better WLB have lower probability of attrition)
- Marital Status (Single employees have a higher attrition probability)
- Environment Satisfaction (Employees with low environment satisfaction have higher attrition probability)
- Total Working Years (Employees with more working years have lower attrition probability)
- Age (Middle-aged 32-46 yr old emplyees have lower attrition probability)
- Years Since Last Promotion (Employees who do not get promoted for a long time have ahigher attrition probability)
- Average Hours (Over worked employees who spend longer hours at work have higher attrition probability)
- Years with current manager (Temporal familiarity with manager lowers attrition probability)
- TrainingTimesLastYear (Employees who are trained less have higher attrition probability)
- NumCompaniesWorked (Employees who have worked at more companies have higher attrition probability)

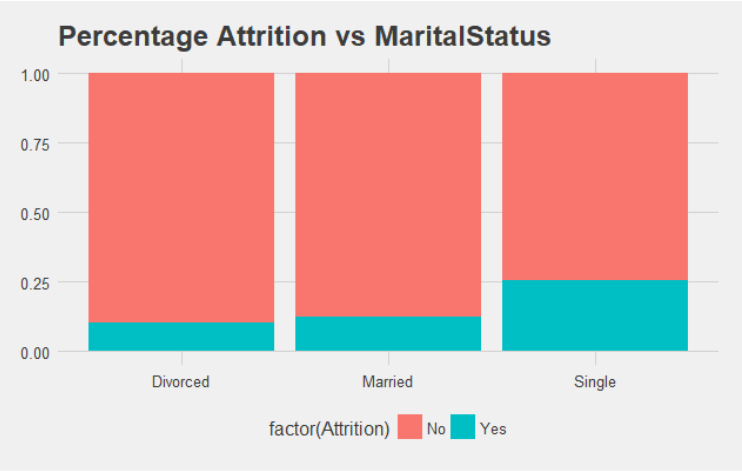
We will investigate this further with Plots in Tableau and R



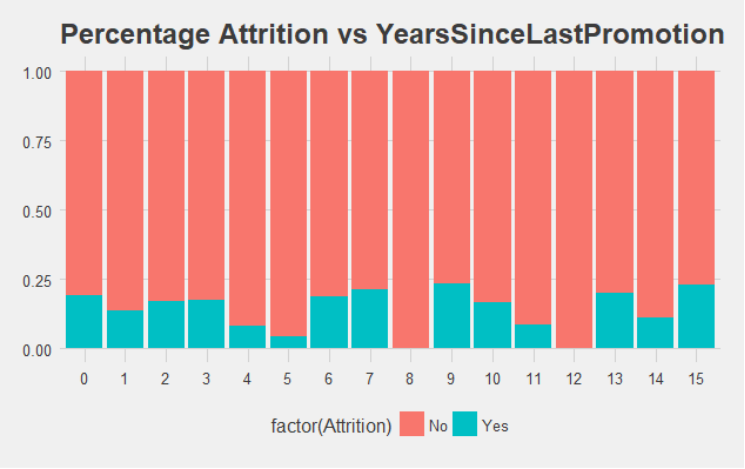
Action : Improve WLB for employees



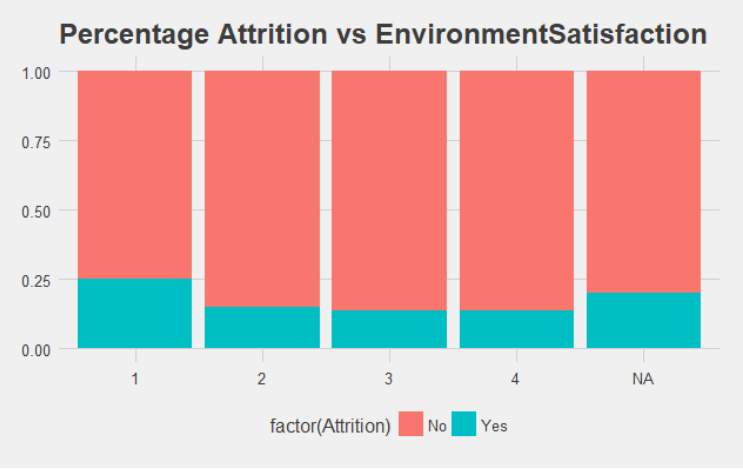
Action : Rotate Employees to spread out travel



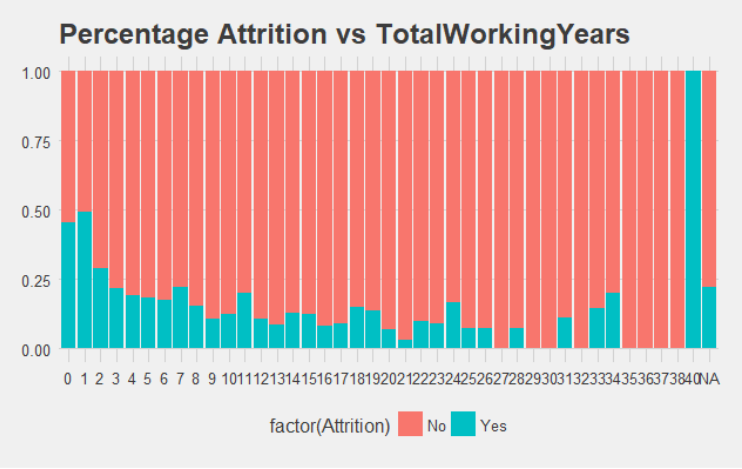
Action : Single people are prone to attrition, try improving other factors such as WLB/Work Environment for them.



Action : Consider employees who are important for promotion especially during milestone years such as 5/10/15 yrs, since milestones are times when employees reflect.



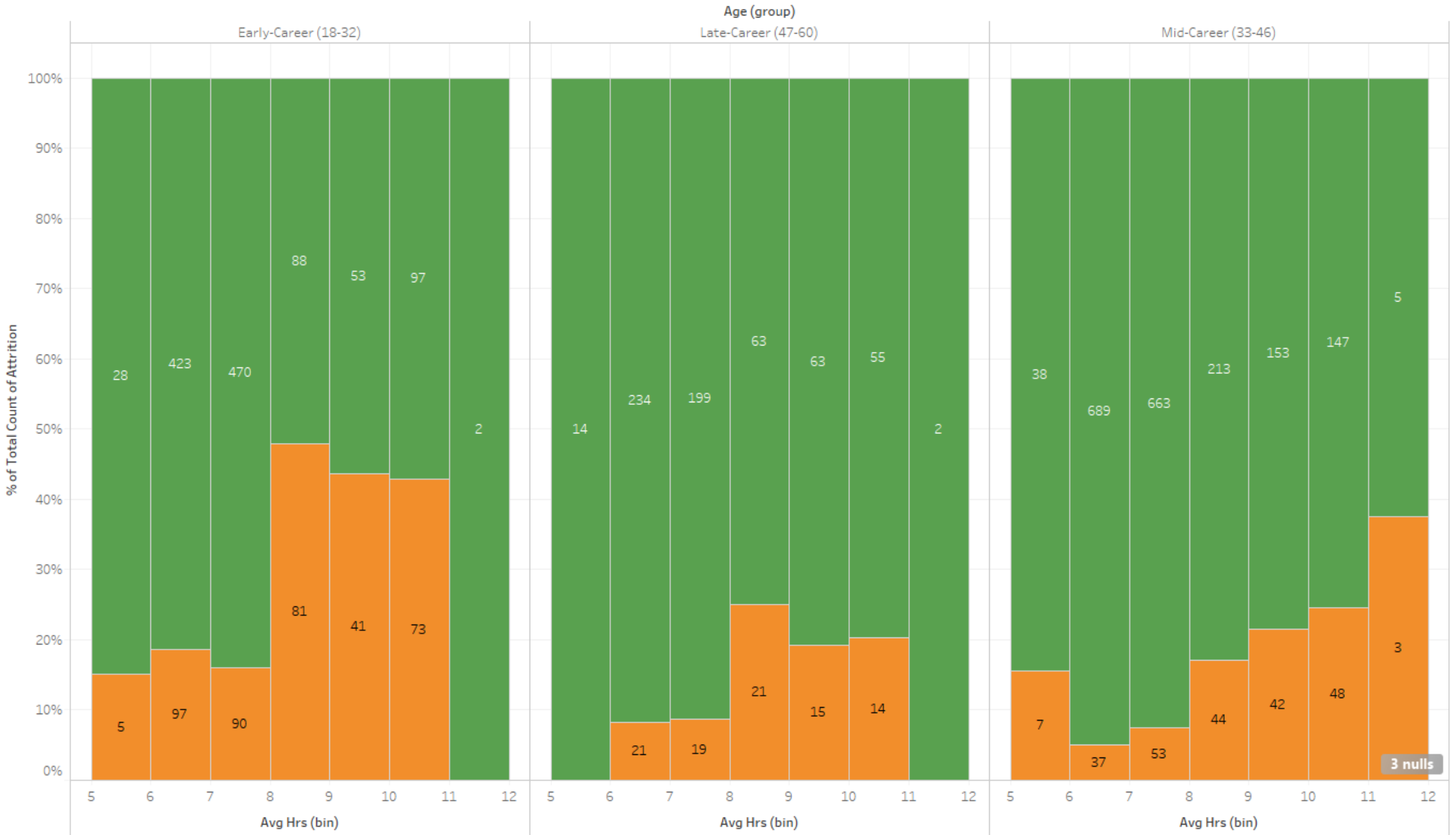
Action : Improve work environment by:
1. Stocking break areas
2. Summer Events/Kids@work day etc.



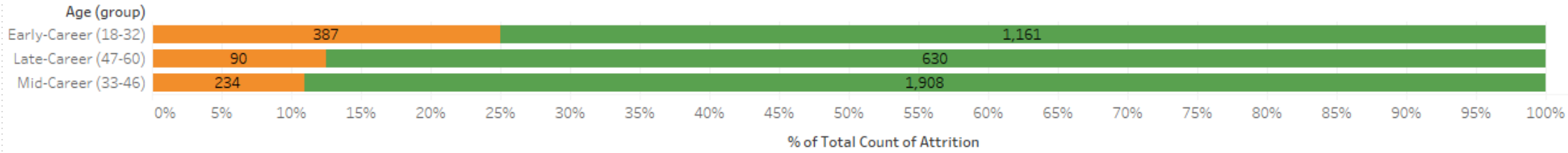
Action : Fresh/young hires are susceptible to attrition. Try and better factors for them.
Proactively train to fill in for people close to retirement

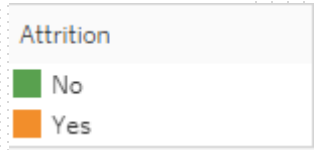

Attribution
No
Yes
Age (group)
Business Travel
Job Role
Early-Career (18-32)
Late-Career (47-60)
Mid-Career (33-46)
Non-Travel
Healthcare Representative 3 18 3 12
Human Resources 6 3 9
Laboratory Technician 6 18 12 33
Manager 15 9 3
Manufacturing Director 3 9 3 3 21
Research Director 3 3 18
Research Scientist 3 24 3 3 63
Sales Executive 39 24 51
Sales Representative 3 9 15
Travel_Frequently
Healthcare Representative 15 21 6 6 3 21
Human Resources 3 6 24
Laboratory Technician 24 21 3 12 12 54
Manager 9 18 3 6 18
Manufacturing Director 15 15 6 6 60
Research Director 6 6 3 6 9 24
Research Scientist 24 42 3 30 12 69
Sales Executive 27 36 6 24 12 72
Sales Representative 3 9 3 6 12
Travel_Rarely
Healthcare Representative 15 69 3 18 12 168
Human Resources 33 6 12 9 45
Laboratory Technician 39 144 9 102 33 255
Manager 18 63 3 51 9 81
Manufacturing Director 9 96 3 60 9 117
Research Director 15 48 3 18 18 60
Research Scientist 63 153 18 87 33 246
Sales Executive 69 183 15 99 36 285
Sales Representative 15 63 33 6 72
0% 20% 40% 60% 80% 100%
% of Total Count of Attrition

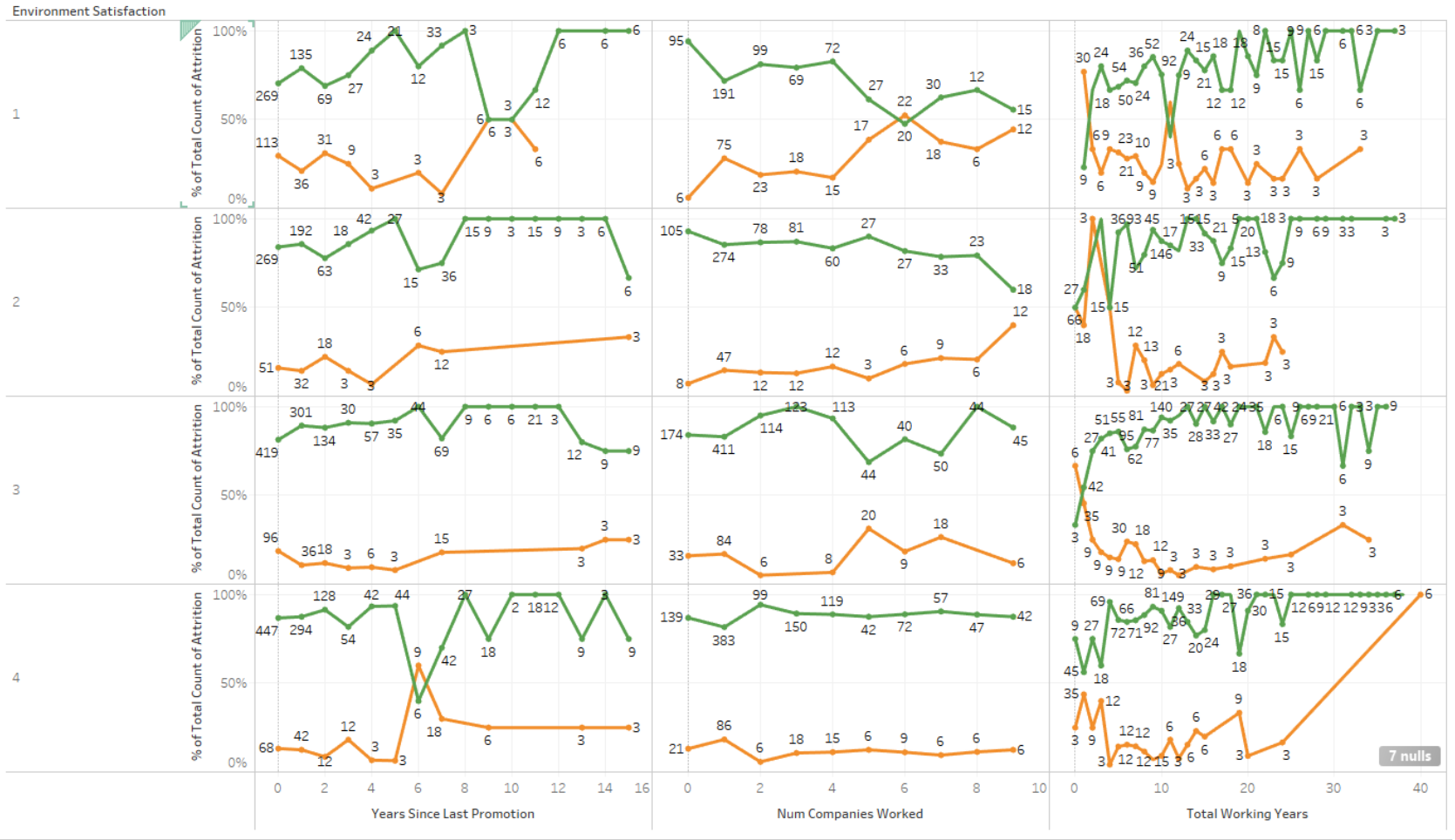
Sheet 5



Attrition
No
Yes


Age (group)
Early-Career (18-32)
Late-Career (47-60)
Mid-Career (33-46)
387
1,161
90
630
234
1,908
0%
5%
10%
15%
20%
25%
30%
35%
40%
45%
50%
55%
60%
65%
70%
75%
80%
85%
90%
95%
100%
% of Total Count of Attrition


Attrition
No
Yes


Environment Satisfaction
1
2
3
4
% of Total Count of Attrition
100%
50%
0%
Years Since Last Promotion
Num Companies Worked
Total Working Years
7 nulls

UpGrad

imt·b

Conclusion

Factors affecting Attrition	Workplace Changes	Immediate Actions
Excessive Business Travel	Rotation	Prevent travel from overloading same employees
Work Life Balance	Allow managers to be lenient when employees need to be away for family/personal reasons.	Instruct managers to encourage their reports to not sacrifice Work life balance.
Marital Status	Not much can be done, people choose to marry or settle down when they feel like.	While hiring new candidates prefer candidates who are not very young everything else being same.
Work Environment	Invest in office supplies/ break areas/ interior upgrades/kids@work days etc.	Upgrade work environment. Office Ergonomics, Refreshments can be stocked up and revisited.
Age/Total Working Years	Mid-Career employees (32-46) are less susceptible to attrition	Younger employees can be appreciated by creating within company awards, at the same time be mindful of employees nearing retirement and make arrangements for replacements proactively.
Years Since Last Promotion	Appreciation/promotion go along way in retaining. Instruct managers to consider employees for Stocks/Pay appraisals if not promotions	Specifically look at employees nearing milestone years 5/10/15 and who have not been promoted for a while.
Average Hours	Could ask managers spread out workload more evenly	Identify employees with high average workhours and ease their workload.
Years with Manager	Encourage managers to build rapport with their reports.	Look at new manager-employee situations for grievances.
Training	Organize more trainings tailored towards people.	Train Younger employees