IBM Attrition Analysis

Apurva Sharma
Shaojie Zhang
Jiaming Guo
Srinivas Reddy Pachika
Wenqi Wu

Whitman
Syracuse University

Agenda

Introduction

Data Description & Visualization

Predicting Variables

Modeling Analysis

Conclusion

IBM

Introduction

Introduction

- American MNC
- Worried HR department!
- 

Description & Visualization

Description & Visualization

Data Overview

Whitman
Syracuse University

Source

- https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset/home

Objective

- Total include 35 variables and 1470 observations

Sample

- a random sample out of the total IBM employees within 3 departments for the previous year this data was collected.

Description & Visualization

Whitman
Syracuse University

Data Dictionary

Variable	Meaning	Levels
Age	Age of the employee	
Attrition	Whether the employee left in the previous year or not	
BusinessTravel	How frequently the employees travelled for business purposes in the last year	
Department	Department in company	
DistanceFromHome	Distance from home in kms	
Education	Education Level	1 'Below College'
		2 'College'
		3 'Bachelor'
		4 'Master'
		5 'Doctor'
EducationField	Field of education	
EmployeeCount	Employee count	
EmployeeNumber	Employee number/id	
EnvironmentSatisfaction	Work Environment Satisfaction Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
Gender	Gender of employee	
JobInvolvement	Job Involvement Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
JobLevel	Job level at company on a scale of 1 to 5	
JobRole	Name of job role in company	

Variable	Meaning	Levels
JobSatisfaction	Job Satisfaction Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
MaritalStatus	Marital status of the employee	
MonthlyIncome	Monthly income in rupees per month	
NumCompaniesWorked	Total number of companies the employee has worked for	
Over18	Whether the employee is above 18 years of age or not	
PercentSalaryHike	Percent salary hike for last year	
PerformanceRating	Performance rating for last year	1 'Low'
		2 'Good'
		3 'Excellent'
		4 'Outstanding'
RelationshipSatisfaction	Relationship satisfaction level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
StandardHours	Standard hours of work for the employee	
StockOptionLevel	Stock option level of the employee	
TotalWorkingYears	Total number of years the employee has worked so far	
TrainingTimesLastYear	Number of times training was conducted for this employee last year	
WorkLifeBalance	Work life balance level	1 'Bad'
		2 'Good'
		3 'Better'
		4 'Best'
YearsAtCompany	Total number of years spent at the company by the employee	
YearsSinceLastPromotion	Number of years since last promotion	
YearsWithCurrManager	Number of years under current manager	

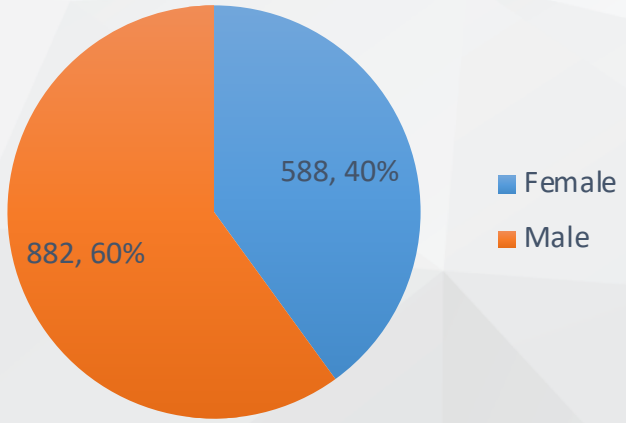
Description & Visualization

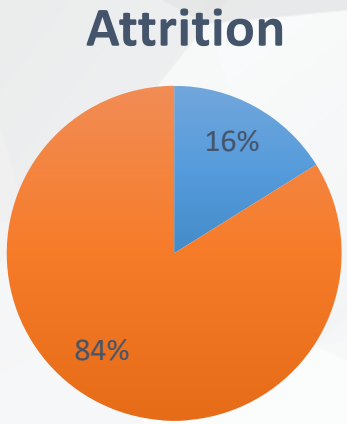
Whitman
Syracuse University

Data Visualization

1. Employee Profile

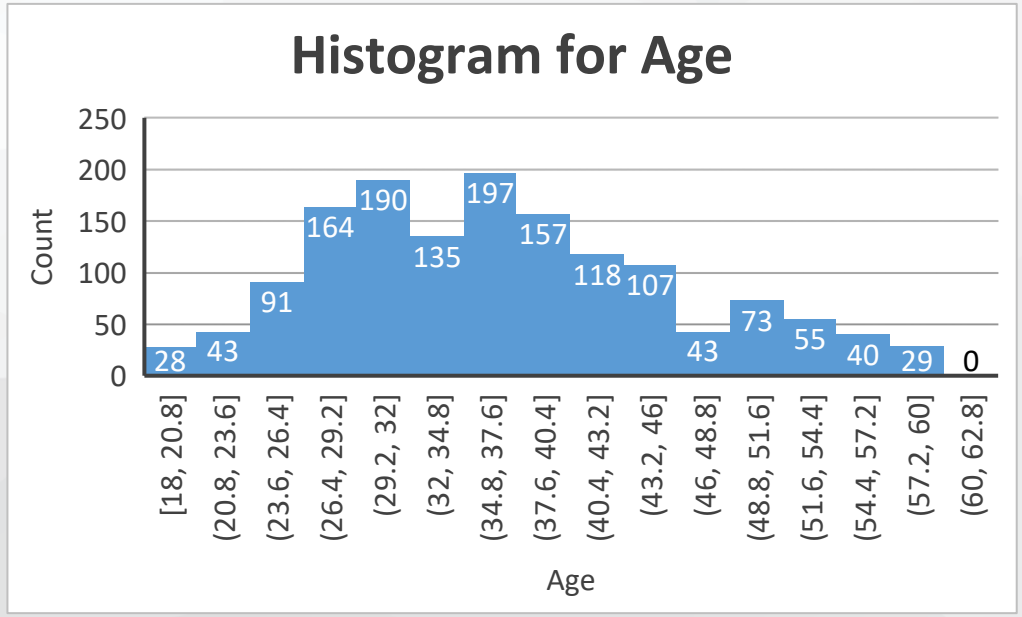
Male>Female





84% Employees choose to stay in IBM

Employee age distribution

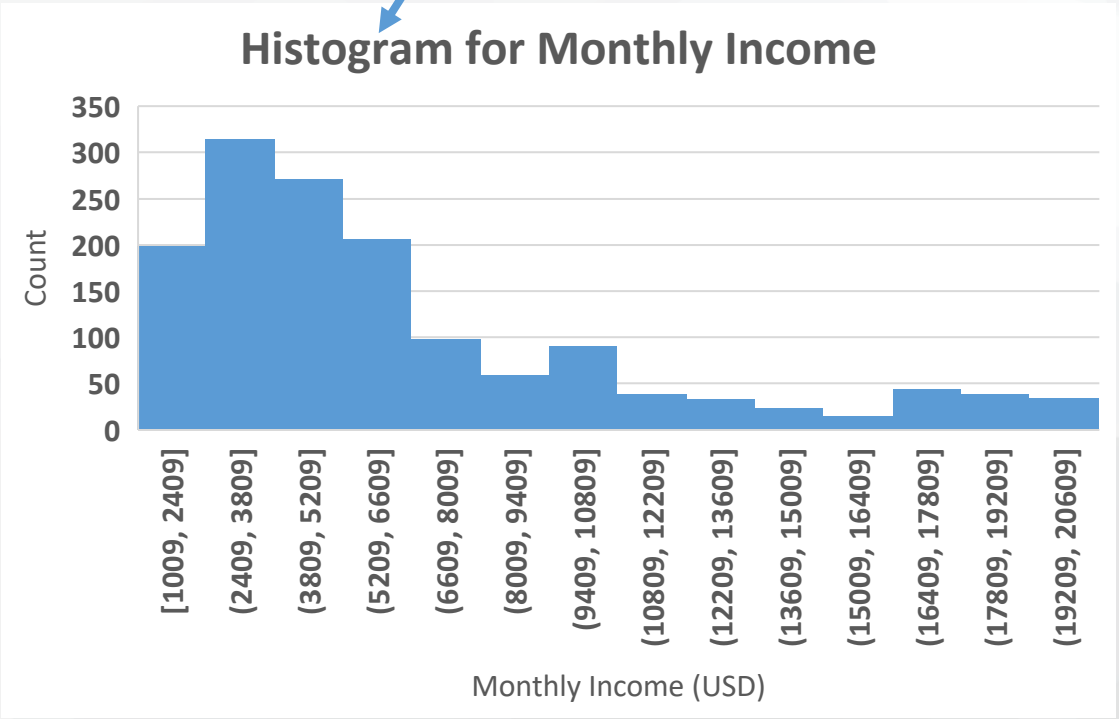


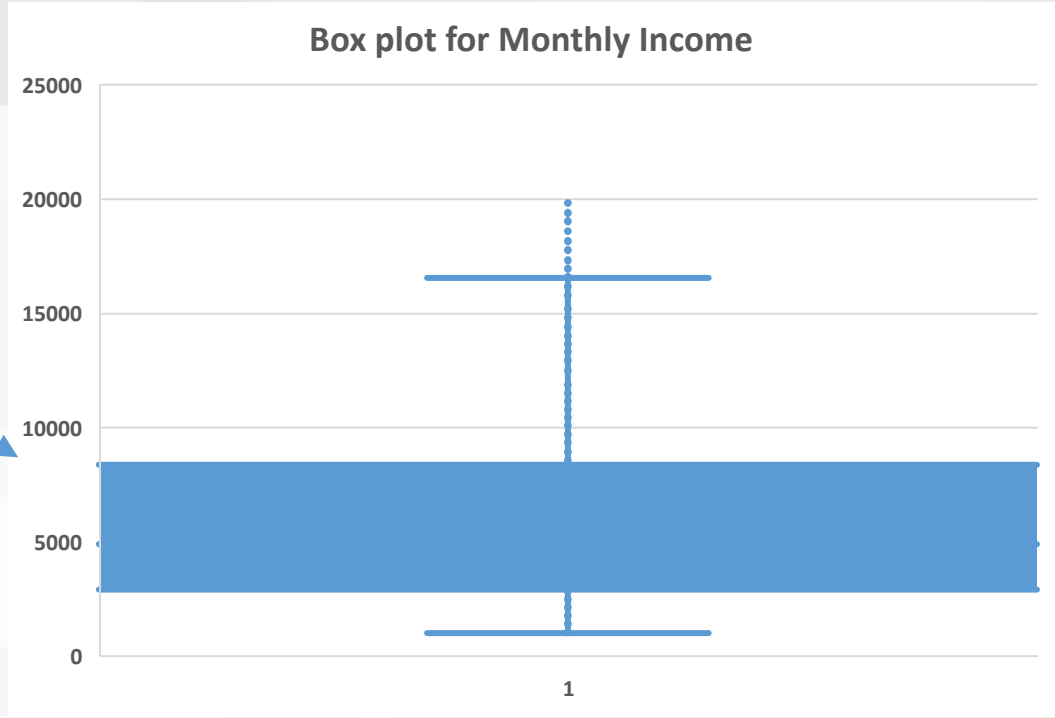
Description & Visualization

2. Monthly Income

2500-8000 USD

Right Skewed

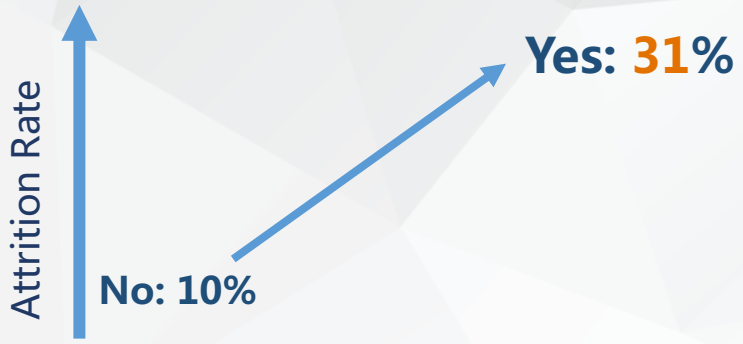


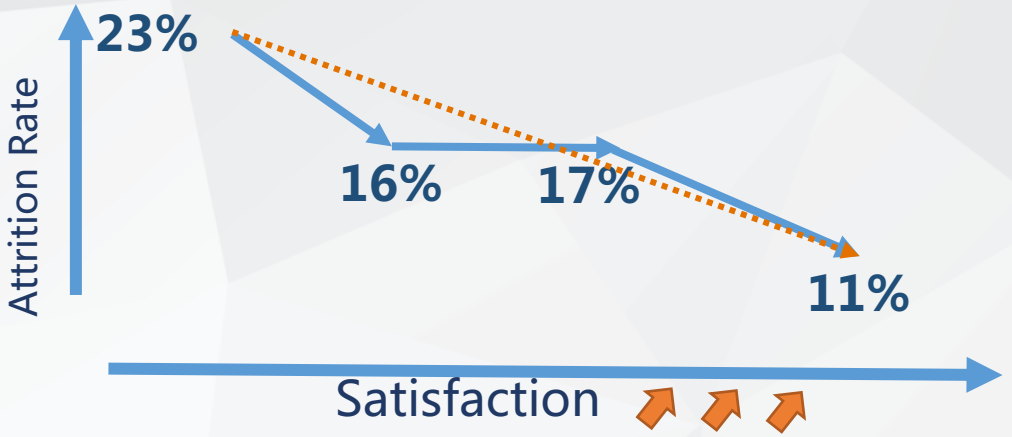


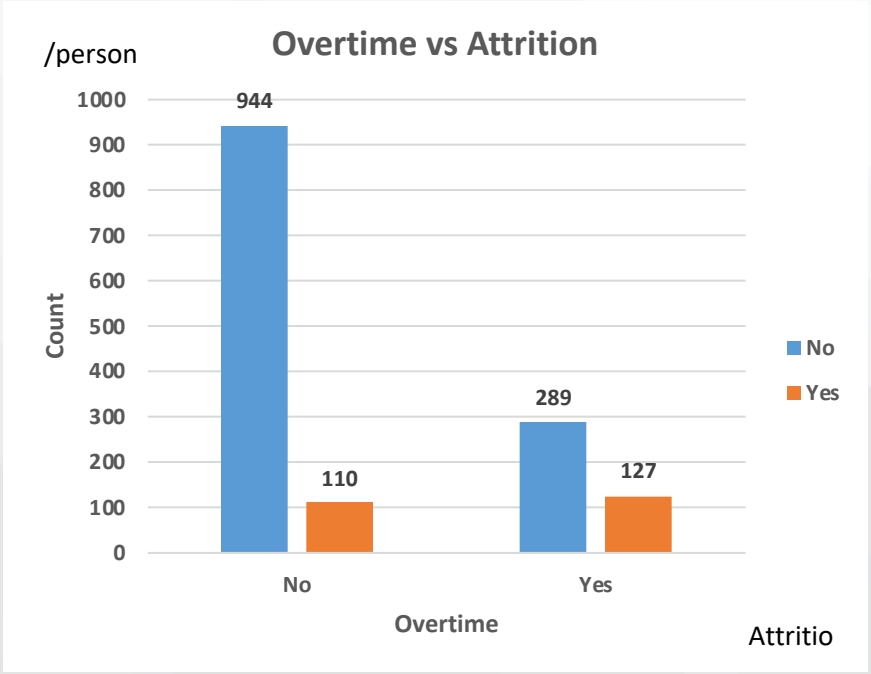
Description & Visualization

Whitman
Syracuse University

3. Cross Group Comparison







Description & Visualization

Winsorization

1%	0	Zscore>3	14
Wins_TotalWorkingYears		Zscore	
99%	35	2<Zscore<=3	68
1%	1	Zscore>3	16
Wins_YearsAtCompany		Zscore	
99%	31	2<Zscore<=3	68
1%	0	Zscore>3	25
Wins_DistanceFromHome		Zscore	
99%	29	2<Zscore<=3	87
1%	1	Zscore>3	0
Wins_Age		Zscore	

Z Score

99% Winsorization

Description & Visualization

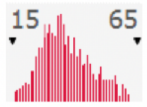
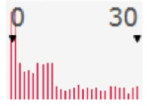
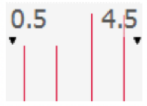
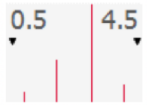
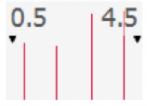
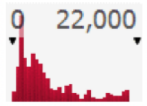
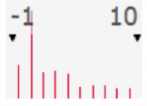
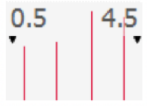
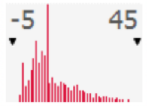
Whitman
Syracuse University

Data Cleaning

- **No NA's , Drop columns (EmployeeNumber,EmployeeCount,StandardHours,Over18 etc.)**
- **Correlation**
- **99% Winsorization on these columns**
  - Age
  - Monthly Income
  - Years with Current Manager
  - Total Working Years
  - Years at Company
  - Distance from Company
  - Years Since Last Promotion

Description & Visualization

Data Summary

Name	Graph	Minimum	Maximum	Mean	Std. Deviation	5%	95%	Count
Age		18	60	36.924	9.135	24	54	1470
DistanceFromHome		1	29	9.193	8.107	1	26	1470
EnvironmentSatisfaction		1	4	2.7218	1.0931	1	4	1470
JobInvolvement		1	4	2.7299	0.7116	1	4	1470
JobSatisfaction		1	4	2.7286	1.1028	1	4	1470
MonthlyIncome		1,009.00	19,999.00	6,502.93	4,707.96	2,097.00	17,856.00	1470
NumCompaniesWorked		0	9	2.6932	2.498	0	8	1470
RelationshipSatisfaction		1	4	2.7122	1.0812	1	4	1470
TotalWorkingYears		0	40	11.28	7.781	1	28	1470

After data cleansing

1470 data points

Conclusion

Whitman
Syracuse University

- Better understanding of the data
- Import factors

Whitman
Syracuse University

Analysis of Association Between Variables

- Study the correlation and the relationship between variables.
- Build model to predict the attrition, and explore interesting and important factors
- Derive useful business insights, and predict the attrition.

IBM

Predicting Variables

- Dependent Variables
- Reasons
- Scatter Plots
- Correlation

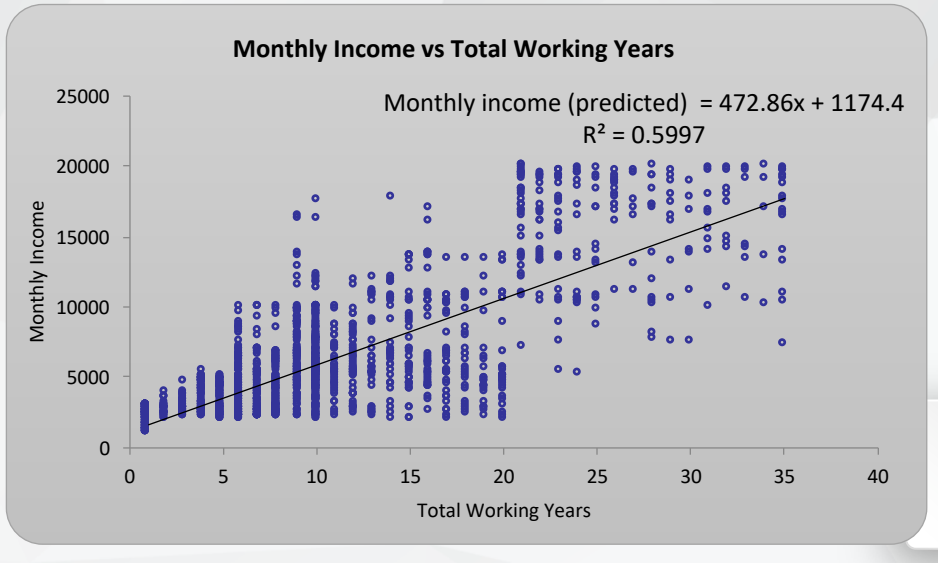
Predicting Variables

Dependent Variables and Reasons

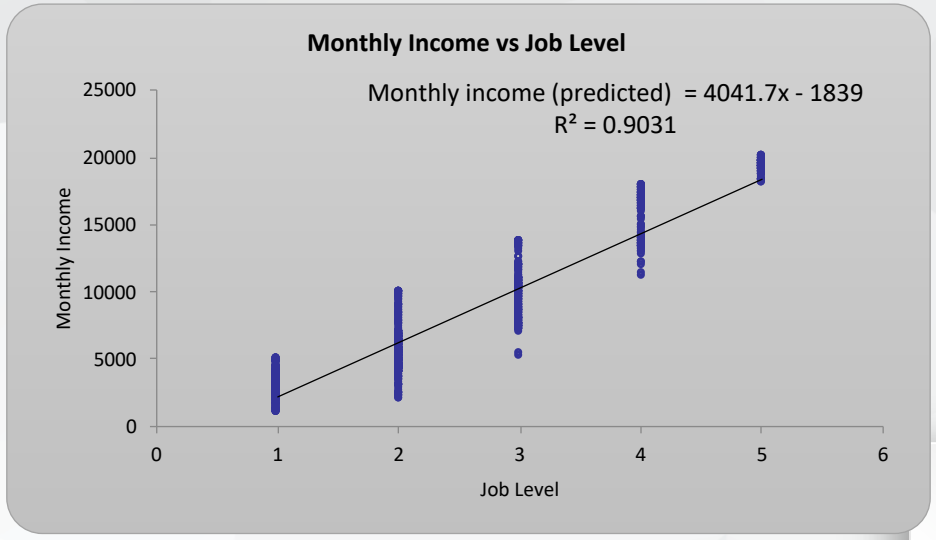
Whitman
Syracuse University

Predicting Variables Scatter Plots

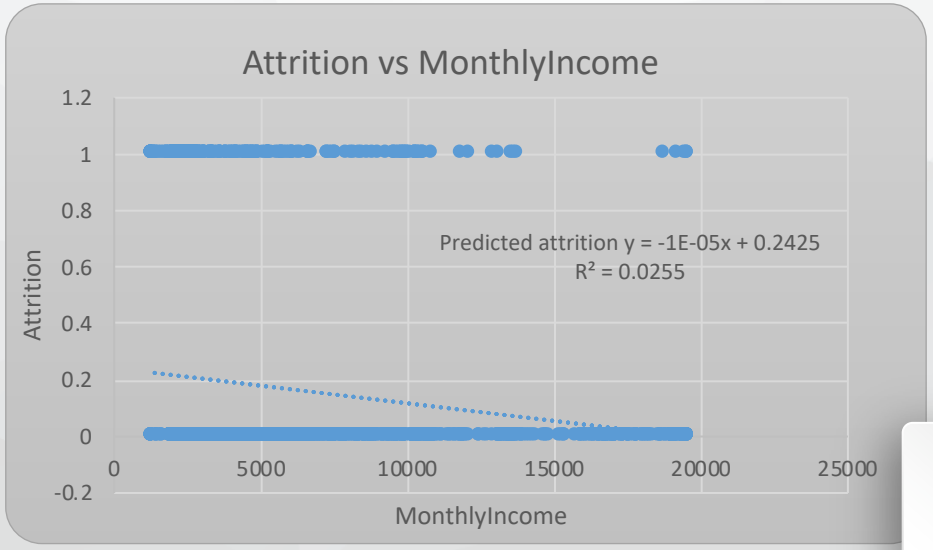
Whitman
Syracuse University



1



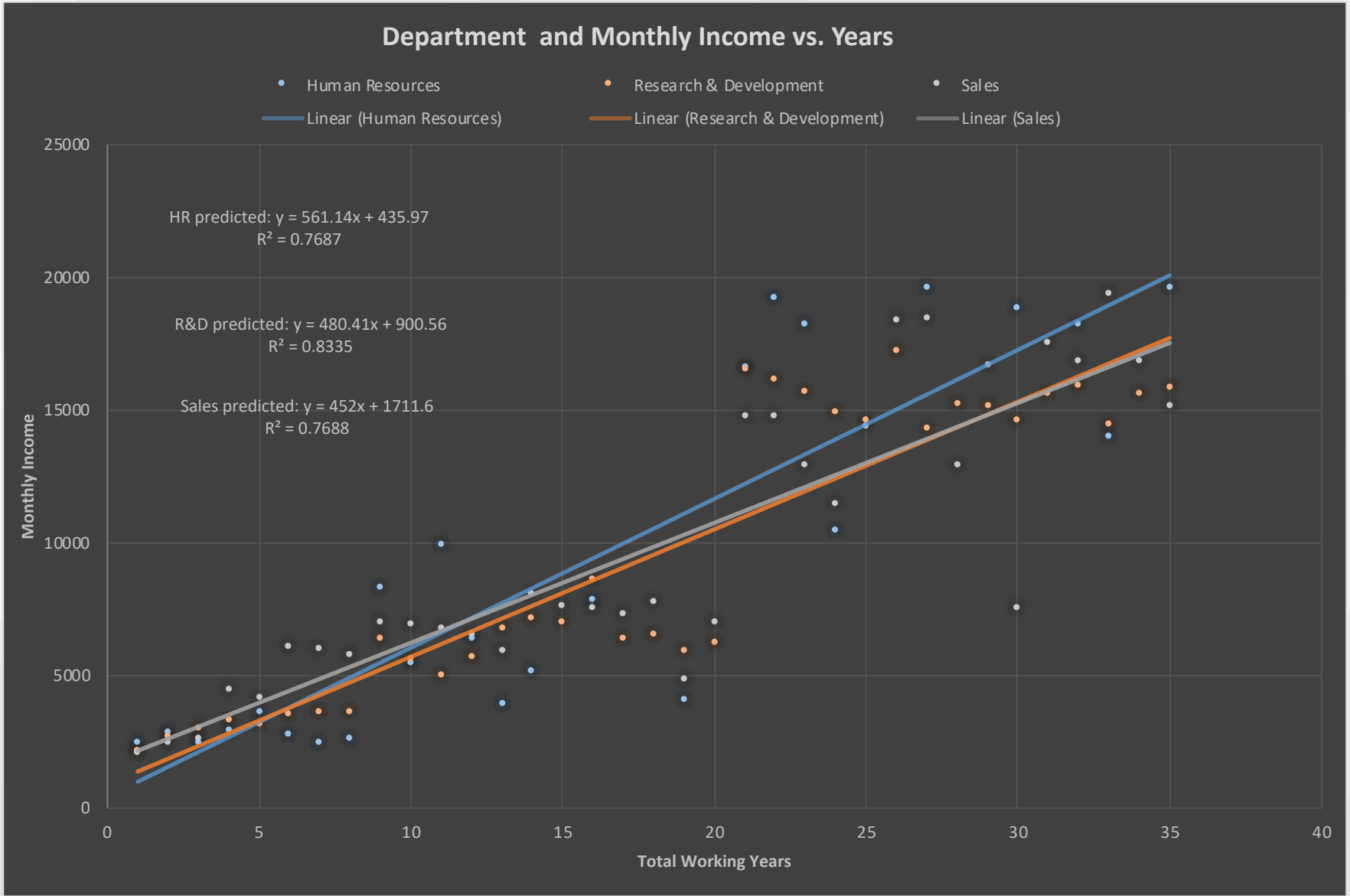
2



3

Predicting Variables Scatter Plots

Whitman
Syracuse University



Predicting Variables

Correlation

Whitman
Syracuse University

	Age	DistanceFromHome
TotalWorkingYears	0.680	0.004
TrainingTimesLastYear	-0.020	-0.037

	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrentMgr
YearsAtCompany	1.000			
YearsInCurrentRole	0.768	1.000		
YearsSinceLastPromotion	0.623	0.548	1.000	
YearsWithCurrentMgr	0.779	0.720	0.511	1.000

	Wins_Age	DistanceFromHom	ironmentSatisfact	HourlyRate	JobInvolvement	JobSatisfaction	ins_MonthlyIncorm	CompaniesWorl	PercentSalaryHike	erformanceRatin	ationshipSatisfact	StockOptionLevel	is_TotalWorkingY	ainingTimesLastYe	WorkLifeBalance	ns_YearsAtCompa	earsInCurrentRol	earsSinceLastPro	_YearsWithCurrentMgr
Wins_Age	1.000																		
Wins_DistanceFromHome	-0.002	1.000																	
EnvironmentSatisfaction	0.011	-0.016	1.000																
HourlyRate	0.024	0.031	-0.050	1.000															
JobInvolvement	0.030	0.009	-0.008	0.043	1.000														
JobSatisfaction	-0.005	-0.004	-0.007	-0.071	-0.021	1.000													
Wins_MonthlyIncome	0.498	-0.017	-0.006	-0.016	-0.015	-0.007	1.000												
NumCompaniesWorked	0.300	-0.029	0.013	0.022	0.015	-0.056	0.149	1.000											
PercentSalaryHike	0.003	0.040	-0.032	-0.009	-0.017	0.020	-0.027	-0.010	1.000										
PerformanceRating	0.002	0.027	-0.030	-0.002	-0.029	0.002	-0.017	-0.014	0.774	1.000									
RelationshipSatisfaction	0.053	0.007	0.008	0.001	0.034	-0.012	0.026	0.053	-0.040	-0.031	1.000								
StockOptionLevel	0.038	0.045	0.003	0.050	0.022	0.011	0.005	0.030	0.008	0.004	-0.046	1.000							
Wins_TotalWorkingYears	0.680	0.004	-0.004	-0.002	-0.006	-0.020	0.774	0.238	-0.021	0.006	0.023	0.011	1.000						
TrainingTimesLastYear	-0.020	-0.037	-0.019	-0.009	-0.015	-0.006	-0.022	-0.066	-0.005	-0.016	0.002	0.011	-0.035	1.000					
WorkLifeBalance	-0.021	-0.027	0.028	-0.005	-0.015	-0.019	0.031	-0.008	-0.003	0.003	0.020	0.004	0.002	0.028	1.000				
Wins_YearsAtCompany	0.309	0.008	-0.001	-0.021	-0.020	-0.005	0.516	-0.117	-0.035	0.004	0.018	0.017	0.624	0.004	0.012	1.000			
YearsInCurrentRole	0.213	0.019	0.018	-0.024	0.009	-0.002	0.363	-0.091	-0.002	0.035	-0.015	0.051	0.460	-0.006	0.050	0.768	1.000		
Win_YearsSinceLastPromotion	0.215	0.010	0.016	-0.027	-0.024	-0.018	0.344	-0.038	-0.023	0.018	0.033	0.015	0.402	-0.001	0.007	0.623	0.548	1.000	
Wins_YearsWithCurrentMgr	0.201	0.012	-0.006	-0.018	0.026	-0.026	0.344	-0.112	-0.011	0.024	-0.002	0.027	0.459	-0.004	0.001	0.779	0.720	0.511	1.000

04 Modeling Analysis

Modeling Analysis

Whitman
Syracuse University

Linear Analysis

Only keep the significant variables

R square: 0.9399

P-Value: <0.0001

Significant Variables:

- Department
- Job Level
- Job Role
- Total working years

Linearly dependent variables ignored: JobRole_Research Dir

Multiple Regression for Wins_MonthlyIncome Summary	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.9697	0.9403	0.9399	1152.631753	0	0

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value
Explained	11	30520500282	2774590935	2088.419811	< 0.0001
Unexplained	1458	1937040418	1328559.957		

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
Constant	26.10347411	177.4689602	0.147087547	0.8831	-322.0182869	374.2252351
Dept_R&D	3080.31505	205.2554182	15.00722893	< 0.0001	2677.687585	3482.942515
Dept_Sales	2469.946269	290.6720259	8.497364896	< 0.0001	1899.766237	3040.126302
JobLevel	2935.831542	65.98099364	44.49510958	< 0.0001	2806.403727	3065.259357
JobRole_Health Rep	-3450.370484	172.1871036	-20.03849541	< 0.0001	-3788.131395	-3112.609573
JobRole_Sales Exe	-2890.389607	273.8458548	-10.55480504	< 0.0001	-3427.563551	-2353.215664
JobRole_Lab	-3839.650109	193.155681	-19.87852539	< 0.0001	-4218.542822	-3460.757396
JobRole_Mgr	907.4443064	185.6641114	4.887559041	< 0.0001	543.2469992	1271.641614
JobRole_Manuf Dir	-3555.740788	169.3881089	-20.99167888	< 0.0001	-3888.011213	-3223.470364
JobRole_Sales Rep	-3242.55461	312.1302521	-10.38846631	< 0.0001	-3854.826936	-2630.282284
JobRole_Research Sci	-3728.131271	193.1882946	-19.29791491	< 0.0001	-4107.087959	-3349.174584
Wins_TotalWorkingYears	43.58883957	6.503756059	6.702102473	< 0.0001	30.83112121	56.34655793

Modeling Analysis

Linear Analysis with interaction variables

R square: 0.9418

Previous R square (0.9399)

P Value: <0.0001

Interaction Variables:

Total Working Year * Department

Only keep the significant variables

Adjusted R-square
0.9399

Linearly dependent variables ignored: JobRole_Research Dir

Multiple Regression for Wins_MonthlyIncome *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.9707	0.9423	0.9418	1133.769996	0	0

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value
Explained	13	30585948209	2352765247	1830.326963	<0.0001
Unexplained	1456	1871592491	1285434.403		

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
Constant	-1092.283601	243.3702584	-4.48815565	<0.0001	-1569.677392	-614.8898107
Dept_R&D	5076.176475	347.4268967	14.61077574	<0.0001	4394.665742	5757.687208
Dept_Sales	3979.090011	476.4006117	8.352403236	<0.0001	3044.585132	4913.59489
JobLevel	2824.097966	67.25704501	41.98962303	<0.0001	2692.166908	2956.029024
JobRole_Health Rep	-3886.750584	180.0725477	-21.58435933	<0.0001	-4239.979925	-3533.521242
JobRole_Sales Exe	-3231.102498	312.4043558	-10.34269349	<0.0001	-3843.913203	-2618.291793
JobRole_Lab	-4488.60727	210.8678368	-21.28635323	<0.0001	-4902.244485	-4074.970055
JobRole_Mgr	508.9453921	193.108327	2.63554348	0.0085	130.1451361	887.7456482
JobRole_Manuf Dir	-4009.687727	178.3895219	-22.47714824	<0.0001	-4359.615654	-3659.7598
JobRole_Sales Rep	-3601.636542	374.4202874	-9.619234483	<0.0001	-4336.097364	-2867.175719
JobRole_Research Sci	-4380.537052	211.0833107	-20.75264519	<0.0001	-4794.596939	-3966.477164
Wins_TotalWorkingYears	166.2075075	19.8013705	8.393737566	<0.0001	127.3652455	205.0497694
Interaction(Wins_TotalWorkingYears,Department = Research & Developr	-134.3605983	19.33262834	-6.949939549	<0.0001	-172.2833781	-96.43781846
Interaction(Wins_TotalWorkingYears,Department = Sales)	-103.6721911	20.81860481	-4.979785723	<0.0001	-144.5098544	-62.83452785

Modeling Analysis

Whitman
Syracuse University

Interpretation on coefficients

Department:	R&D
Job Role:	Manager
Job Level:	3

-1092.28+
5076.18* R&D Department+
2824.09* JobLevel(3)+
508.94* JobRole_Mgr+
-134.36* TotalWorkingYear* R&D
= 12830.77

Department:	Sales
Job Role:	Manager
Job Level:	3

-1092.28+
3979.09* Sales Department+
2824.09* JobLevel(3)+
508.94* JobRole_Mgr+
-103.67* TotalWorkingYear* Sale
= 11764.37

Department:	Sales
Job Role:	Manager
Job Level:	4

-1092.28+
3979.09* Sales Department+
2824.09* JobLevel(4)+
508.94* JobRole_Mgr+
-103.67* TotalWorkingYear* Sale
= 14588.47

Modeling Analysis

Whitman
Syracuse University

Hypothesis Test

Assumption:

Population Mean (μ) = Sample Mean ($\overline{X}$)

Hypothesis:

H0: $\mu \leq 6300$

Ha: $\mu > 6300$

For:

α =5%

α =1%

	Monthly Income
Hypothesis Test (One-Sample)	
Sample Size	1470
Sample Mean	6502.93
Sample Std Dev	4707.96
Hypothesized Mean	6300
Alternative Hypothesis	> 6300
Standard Error of Mean	122.79
Degrees of Freedom	1469
t-Test Statistic	1.6526
p-Value	0.0493
Null Hypoth. at 10% Significance	Reject
Null Hypoth. at 5% Significance	Reject
Null Hypoth. at 1% Significance	Don't Reject

Modeling Analysis

Whitman
Syracuse University

Logistic Regression

P-Value: <0.0001

Significant Variables:

- Age
- Business Travel
- Distance from Home
- Environment Satisfaction
- Gender
- Job Involvement
- Marital Status
- Monthly Income
- Number of Company Worked
- Overtime
- Relationship Satisfaction
- Training Time Last Year
- Work Life Balance
- Years In Current Role
- Years Since Last Promotion

Logistic Regression for Attrition_yes

Summary Measures

Null Deviance	1298.582701
Model Deviance	936.706791
Improvement	361.8759098
p-Value	< 0.0001

Regression Coefficients	Coefficient	Standard Error	Wald Value	p-Value	Lower Limit	Upper Limit	Exp(Coef)
Constant	3.043755875	0.787084788	3.867125783	0.0001	1.501069692	4.586442059	20.98390836
Wins_Age	-0.046146781	0.01154707	-3.996405908	< 0.0001	-0.068779039	-0.023514522	0.954901791
BusinessTravel_Frequently	1.842739602	0.396001365	4.653366792	< 0.0001	1.066576927	2.618902278	6.313811923
BusinessTravel_rarely	1.026761106	0.369621608	2.777870897	0.0055	0.302302756	1.751219457	2.792008157
Wins_DistanceFromHome	0.038782414	0.010025783	3.868267824	0.0001	0.019131879	0.058432949	1.039544269
EnvironmentSatisfaction	-0.387582006	0.077230964	-5.018479435	< 0.0001	-0.538954695	-0.236209316	0.678695975
Female	-0.395973129	0.174577105	-2.268184766	0.0233	-0.738144255	-0.053802004	0.67302478
JobInvolvement	-0.574442735	0.11540386	-4.977673503	< 0.0001	-0.8006343	-0.34825117	0.563018532
JobSatisfaction	-0.396756376	0.076335038	-5.19756574	< 0.0001	-0.546373049	-0.247139702	0.672497842
Marital_Single	1.356658274	0.247244373	5.48711486	< 0.0001	0.872059303	1.841257245	3.883195022
Marital_Married	0.357782096	0.2455358	1.457148392	0.1451	-0.123468072	0.839032264	1.43015395
Wins_MonthlyIncome	-0.000113362	2.89145E-05	-3.920590633	< 0.0001	-0.000170034	-5.66894E-05	0.999886645
NumCompaniesWorked	0.151906564	0.034539205	4.398090946	< 0.0001	0.084209722	0.219603405	1.164051467
Overtime_Yes	1.768587878	0.178611482	9.901871139	< 0.0001	1.418509373	2.118666382	5.86256885
RelationshipSatisfaction	-0.227461378	0.077820507	-2.922897679	0.0035	-0.379989571	-0.074933185	0.796553186
TrainingTimesLastYear	-0.154114378	0.06884451	-2.238586314	0.0252	-0.289049618	-0.019179138	0.857173974
WorkLifeBalance	-0.265252602	0.115569086	-2.29518646	0.0217	-0.491768011	-0.038737193	0.767012177
YearsInCurrentRole	-0.165141084	0.034376884	-4.803841036	< 0.0001	-0.232519776	-0.097762392	0.847774088
Wins_YearsSinceLastPromotion	0.168078311	0.035261827	4.766579726	< 0.0001	0.09896513	0.237191493	1.183029251

Modeling Analysis

Interpretation on coefficients

Whitman
Syracuse University

Constant	Age	BusinessTravel_Frequently	BusinessTravel_rarely	DistanceFromHome	EnvironmentSatisfaction	Female	JobInvolvement	JobSatisfaction	Marital_Single	Marital_Married
3.04	-0.05	1.84	1.03	0.04	-0.39	-0.40	-0.57	-0.40	1.36	0.36
1	36	1	0	7	3	0	3	3	1	0
1	36	1	0	7	3	0	3	3	1	0

MonthlyIncome	NumCompaniesWorked	Overtime_Yes	RelationshipSatisfaction	TrainingTimesLastYear	WorkLifeBalance	YearsInCurrentRole	YearsSinceLastPromotion
0.00	0.15	1.77	-0.23	-0.15	-0.27	-0.17	0.17
4919	2	1	3	3	3	3	1
4919	2	0	3	3	3	3	1

Prediction

	Ln(Z)	Z
	P(Arrition=Yes)	
Overtime=1	50.60%	0.0239
Overtime=0	14.87%	-1.7446
	35.73%	

Conclusion

Conclusion

Whitman
Syracuse University

- Who may benefit from our finding: Senior HR Management Department in IBM and other similar tech-companies
- Business Insight
  - How to control the attrition rate
    - do not subject their employees to increased ‘overtime’ and ‘ business travel’
    - to retain their employees , since ‘monthly salary’ positively drives the attrition rate significantly , the company must consider the employees’ total work experience, their department
    - company must invest on their employees in terms of vocational trainings offered over the year and focus on their job involvement
    - If the employee has worked for a high number of companies in their past career, the employee might prefer to change their firm every few years and this might have nothing to do with IBM