The Future of Work

Retention and Compensation Analysis

Analysis Presented to Steven Williams, CEO of Pepsico NA and

Patrick McLaughlin, CHRO of Frito-Lay

Laura Lazarescou, Data Scientist

SMU MS in Data Science Program

Introduction

- Context and Data Overview
- Exploratory Data Analysis for Two Models
- Model Overview
  - How was the model developed?
  - How accurate is the model?
  - Attrition trends and insights
- Conclusion and Recommendations

Context and Data Overview

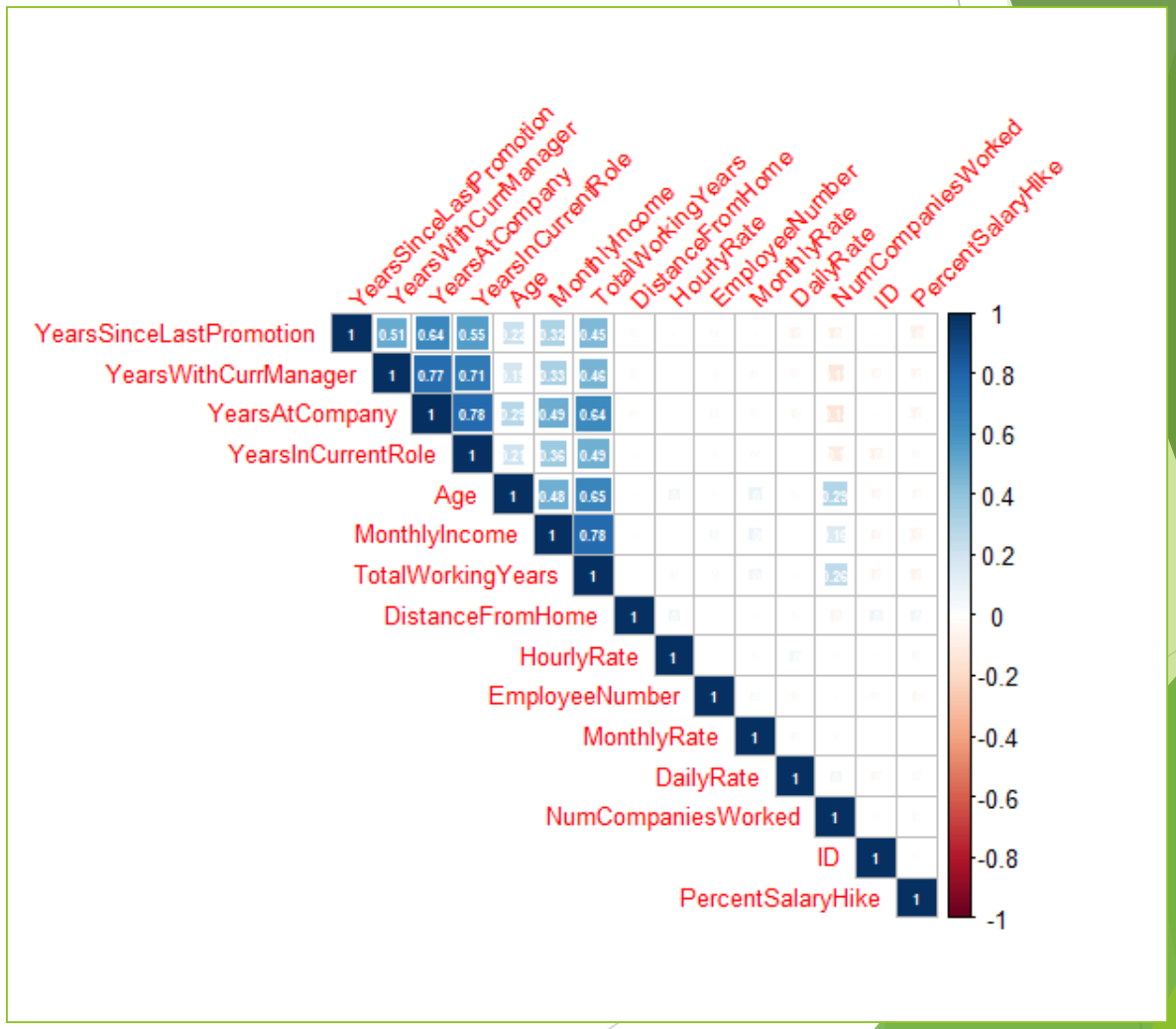
- Frito-Lay seeks to transform its culture and reduce employee turnover. With an evolving demographic and low unemployment rate in North America, it is imperative that Frito-Lay become a “Best Place to Work” in order to manage operating costs and maintain market leadership.
- Data Overview – Three (3) datasets
  1. Train dataset – master with all data needed to create (2) models
     - 870 anonymized employee data records
     - 36 factors
  2. Test dataset for Attrition
  3. Test dataset for Salary
- DDSAnalytics has developed two models
  - Attrition: Identify key factors that lead to employee attrition
  - Salary: Understand the role that salary plays in retaining employees

Exploratory Data Analysis (EDA) Salary and Attrition Models

- Rstudio and R compatibility
  - Converted numeric categorical factors to as.factor
  - NA grooming – no NAs
- Evaluate data compliance with statistical assumptions
  - Normal distribution
  - Equal variance
  - Independence
  - Outliers, Leverage
- Data value analysis
  - Train: No NA or null values.
  - Single value factors: Over18, StandardHours, EmployeeCount removed

Collinearity: Identify and Reduce Redundant Variables

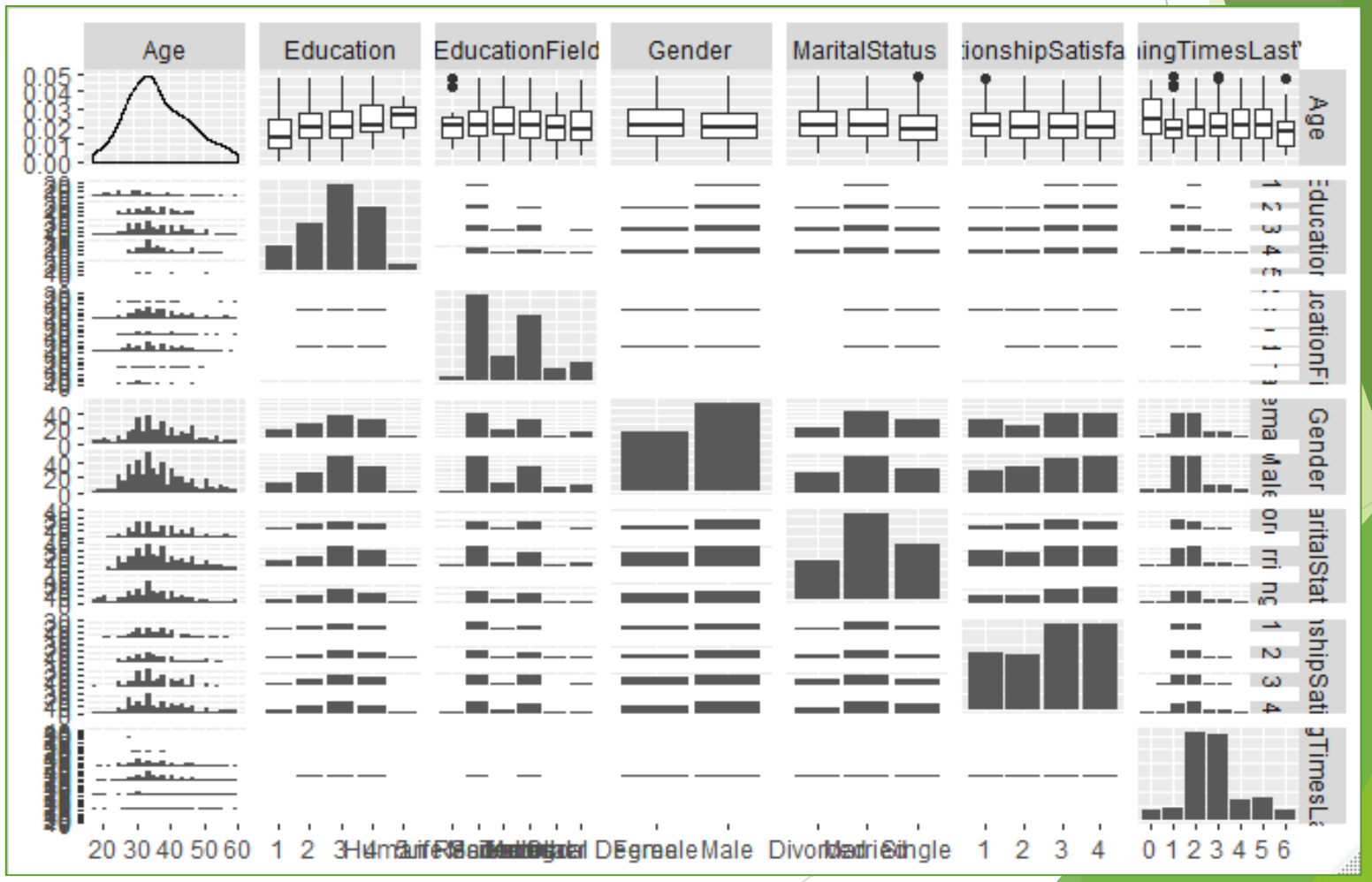
- Few numeric variables are highly correlated.
- Alternate selection of
  - YearsAtCompany
  - YearsWithCurrentManager
  - YearsInCurrentRole
  - TotalWorkingYears



Evaluate Statistical Assumptions People-Focused Variables (1/4)

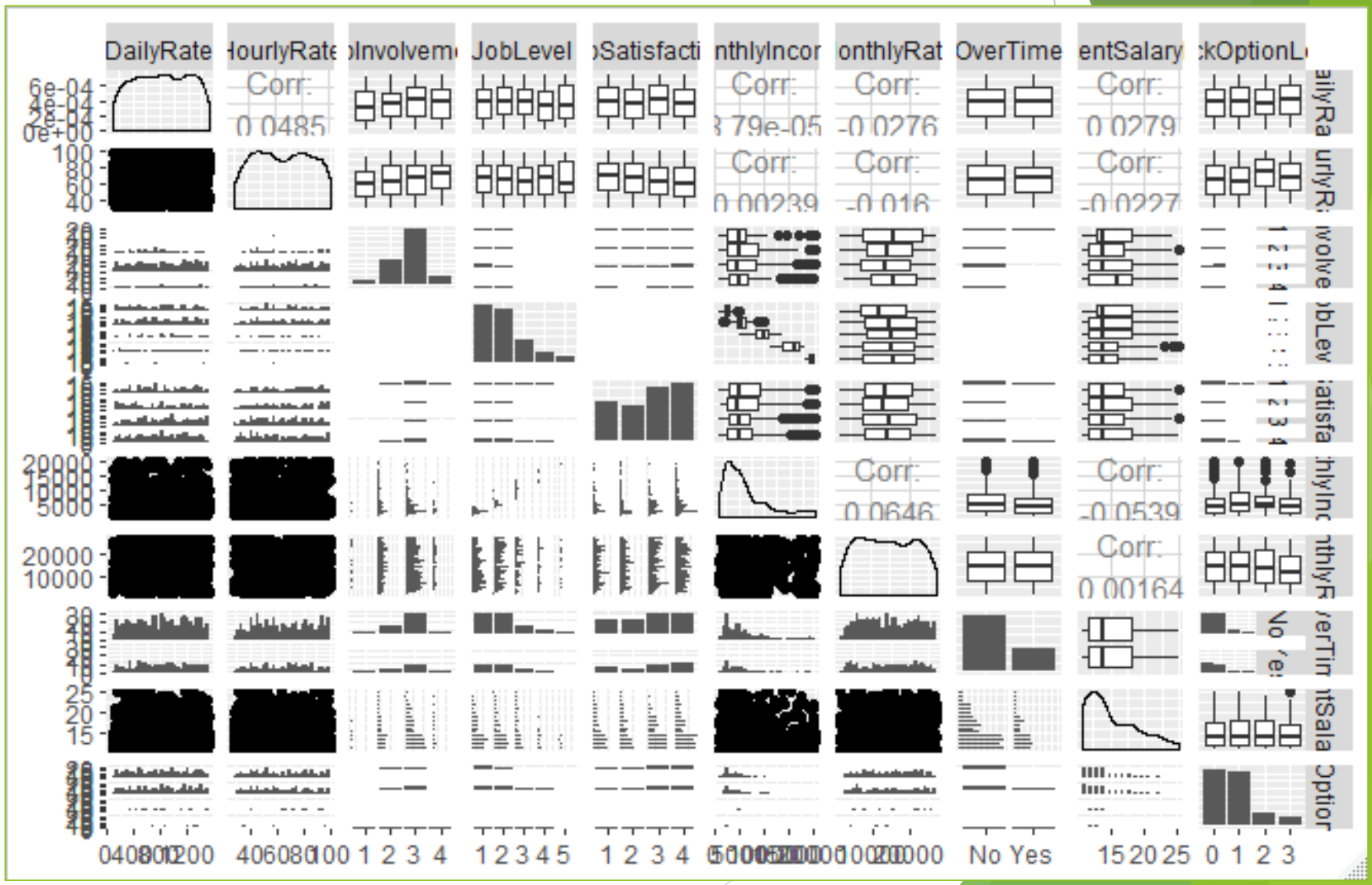
- Normal Distribution
- Equal Variance
- Independence
- Outliers
- Leverage

- Age
- Education
- Education Field
- Gender
- Marital Status
- Relationship Satisf.
- # of Training Times since last year



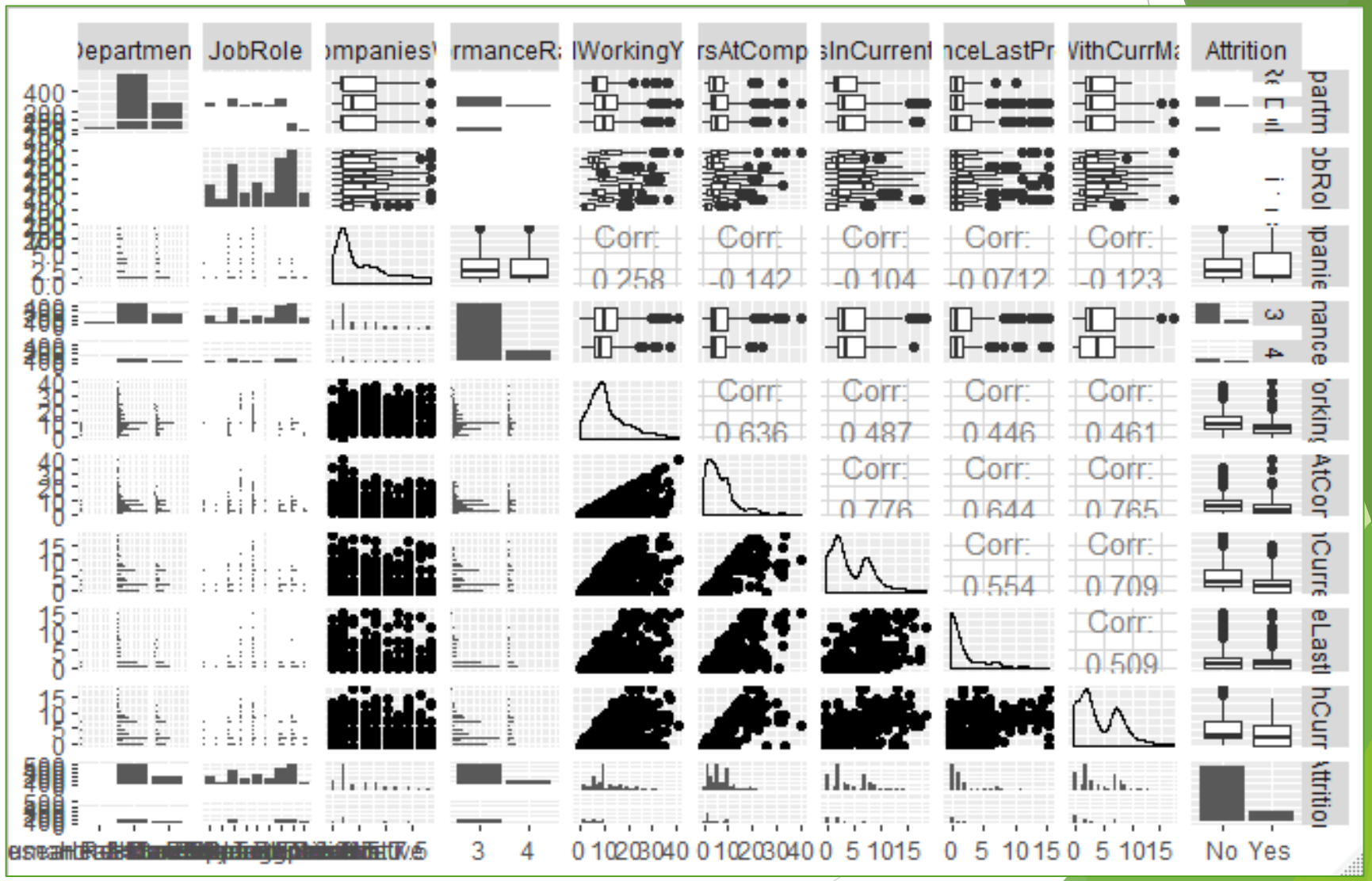
Income-Oriented Variables (2/4)

- Daily Rate
- Hourly Rate
- Job Involvement
- Job Level
- Job Satisfaction
- Monthly Income
- Monthly Rate
- Overtime
- % Salary Hike
- Stock Option Level



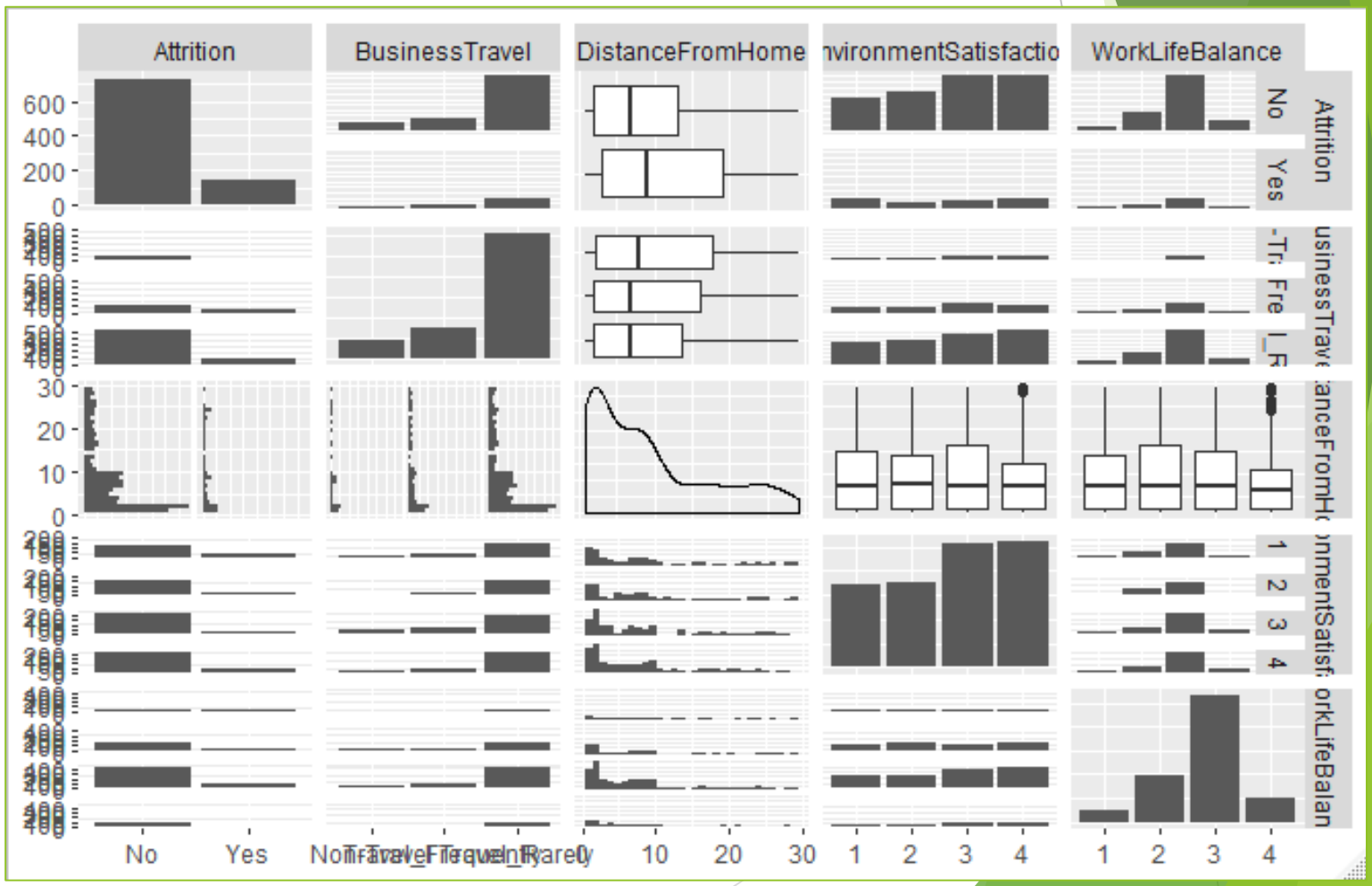
Organizational Variables (3/4)

- Department
- JobRole
- # of Companies
- Performance
- Total Working Yrs
- Years at Company
- Years in Role
- Years Since Promotion
- Years w/Manager



Work-Life Balance Variables

- Attrition
- Business Travel
- Distance from Home
- Environmental Satisfaction
- Work-Life Balance



Salary Model: Multiple Linear Regression

- Forward, Backward, Stepwise approach with full dataset
  - Iterative approach to maximum number of factors
  - Top three factors (Frito-Lay request)
  - Differences in R-squared and CV PRESS (show table of different values)
- Preferred model – Generate Predictions
- Predictive accuracy and sensitivity

Salary Regression Analysis

We can model Frito-Lay salaries with 95.3% accuracy using a Custom Linear Regression Model.

Regression Model	# of Variables	R-squared
Forward	68	.943
Backward	61	.942
Stepwise	56	.948
Custom	20	.953

```
> step.modelF$results
  nvmax     RMSE  Rsquared      MAE  RMSESD RsquaredSD    MAESD
1    68 1047.102 0.9432461 804.358 155.628 0.03472316 90.96424
> step.modelF$bestTune
```

```
> step.modelB$results
  nvmax     RMSE  Rsquared      MAE   RMSESD RsquaredSD    MAESD
1    61 1042.673 0.9423035 805.4494 144.2418 0.03669518 89.39478
```

```
> step.models$results
  nvmax     RMSE  Rsquared      MAE   RMSESD RsquaredSD    MAESD
1    56 1044.788 0.9475428 802.1414 118.9361 0.01404599 86.64594
```

```
Residual standard error: 1004 on 849 degrees of freedom
Multiple R-squared:  0.9534,    Adjusted R-squared:  0.9523
F-statistic: 868.6 on 20 and 849 DF,  p-value: < 2.2e-16
```

Salary Regression – Top Factors

```
Call:
lm(formula = MonthlyIncome ~ ., data = SalaryCS2)

Residuals:
    Min      1Q  Median      3Q     Max
-3192.4  -624.0  -107.2   591.8  4290.2

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     3242.7365   263.0432  12.328  < 2e-16 ***
Age                                0.3112     5.0792   0.061  0.95116
GenderMale                        90.0442    70.1184   1.284  0.19943
JobLevel2                       1720.3574   140.1991  12.271  < 2e-16 ***
JobLevel3                       4944.4599   187.6762  26.346  < 2e-16 ***
JobLevel4                       8264.7602   283.3652  29.166  < 2e-16 ***
JobLevel5                      10969.7570   333.1016  32.932  < 2e-16 ***
JobRoleHuman Resources         -1113.3858   252.1711  -4.415 1.14e-05 ***
JobRoleLaboratory Technician   -1237.6215   175.3381  -7.058 3.50e-12 ***
JobRoleManager                  3340.5463   238.0458  14.033  < 2e-16 ***
JobRoleManufacturing Director    112.7413   158.4925   0.711  0.47707
JobRoleResearch Director        3489.3612   211.7182  16.481  < 2e-16 ***
JobRoleResearch Scientist      -1028.7346   178.7720  -5.754 1.21e-08 ***
JobRoleSales Executive           -16.3693   136.6264  -0.120  0.90466
JobRoleSales Representative    -1232.3453   222.0804  -5.549 3.84e-08 ***
BusinessTravelTravel_Frequently  194.1363   132.6383   1.464  0.14366
BusinessTravelTravel_Rarely      335.5779   112.2113   2.991  0.00286 **
StockOptionLevel1                 63.5591    74.9736   0.848  0.39681
StockOptionLevel2                 15.6125   124.7821   0.125  0.90046
StockOptionLevel3                -27.7217   145.7642  -0.190  0.84921
TotalWorkingYears                 45.4638     8.9254   5.094 4.33e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1004 on 849 degrees of freedom
Multiple R-squared:  0.9534,    Adjusted R-squared:  0.9523
F-statistic: 868.6 on 20 and 849 DF,  p-value: < 2.2e-16
```

Strongest Contributors
Job Level
Job Role: Some job roles may be underpaid vs. market.
Business Travel: Surprising relationship. Higher travel has lower pay. Recommend review and compare with Attrition.
Gender: Evaluate on a per-role basis. Other factors held constant, males earn $90 more than females.

Attrition Model: Classification

- Attrition is a Yes/No categorical value. No detail on voluntary or involuntary
- 16.1% Attrition: 140 employees / 870 Total
- Both models are Naïve Bayes formulas
  - Model 1: 78.5% Accuracy / 83.19% Sensitivity / 48.57% Specificity – 32 variables
  - Model 2: 79.31% Accuracy / 82.8% Sensitivity / 63.04% Specificity – 14 variables

```
     No Yes
No   188  18
Yes   38  17

               Accuracy : 0.7854
                 95% CI : (0.7306, 0.8337)
    No Information Rate : 0.8659
    P-Value [Acc > NIR] : 0.99987

                  Kappa : 0.2558

 Mcnemar's Test P-Value : 0.01112

            Sensitivity : 0.8319
            Specificity : 0.4857
         Pos Pred Value : 0.9126
         Neg Pred Value : 0.3091
             Prevalence : 0.8659
         Detection Rate : 0.7203
   Detection Prevalence : 0.7893
      Balanced Accuracy : 0.6588

       'Positive' Class : No
```

```
     No Yes
No   178  17
Yes   37  29

               Accuracy : 0.7931
                 95% CI : (0.7388, 0.8406)
    No Information Rate : 0.8238
    P-Value [Acc > NIR] : 0.914152

                  Kappa : 0.3915

 Mcnemar's Test P-Value : 0.009722

            Sensitivity : 0.8279
            Specificity : 0.6304
         Pos Pred Value : 0.9128
         Neg Pred Value : 0.4394
             Prevalence : 0.8238
         Detection Rate : 0.6820
   Detection Prevalence : 0.7471
      Balanced Accuracy : 0.7292

       'Positive' Class : No
```

- Age
- Department
- MonthlyIncome
- MaritalStatus
- BusinessTravel
- DistanceFromHome
- EnvironmentSatisfaction
- JobInvolvement
- JobLevel
- OverTime
- PerformanceRating
- StockOptionLevel
- TrainingTimesLastYear
- WorkLifeBalance

Conclusions

- Data was very clean but extensive and largely categorical.
- Reduction in number of variables improved accuracy, sensitivity and specificity
- Salary is relatively predictable: 95% accuracy with few variables
- Attrition is more evasive
  - Need clarification if all Attrition was voluntary?
  - Mean salary for Attrition=Yes is lower than Attrition = No. Consider market-level analysis on salary.
  - Work-Life Balance is more important to younger workers. Evaluate travel requirements with pay scale. Consider remote work options.
- Next Steps: Continue analysis of role-based attrition and potential gender gap.