Human Resource Analytics
Exploratory Analysis

Ryan Fox, rfox2@bellarmine.edu
Bailey Korfhage, bkorfhage2@bellarmine.edu

I. INTRODUCTION

Our Chosen data set contains an employer's information about their employees. The data relates their personal lives at home to their lives at work. Such as their happiness in relationships and their performance at work. We chose this data set because we thought it would be interesting to see if there was a strong connection between work and home life.

II. DATA SET DESCRIPTION

Our data set contains 1470 observations with 35 variables. This software did not contain its own description of data type for each variable. For this dataset we had no null data. A listing of the variables we used is shown in **Table 1**.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
Age	Ratio	0%
Business travel	Nominal	0%
Department	Nominal	0%
Distance from home	Ratio	0%
Employee field	Nominal	0%
Employee number	Interval	0%
Environment Satisfaction	Interval	0%
Gender	Nominal	0%
Hourly rate	Ratio	0%
Job involvement	Interval	0%
Job level	Interval	0%
Job role	Nominal	0%
Job satisfaction	Interval	0%
Marital Status	Nominal	0%
Monthly income	Ratio	0%
Monthly rate	Ratio	0%
Number of companies worked	Ratio	0%
Over 18	Binary	0%
Overtime	Binary	0%
Percent salary hike	Ratio	0%
Performance rating	Interval	0%
Total working years	Ratio	0%
Training times last year	Ratio	0%
Years at company	Ratio	0%
Years in current role	Ratio	0%
Years since last promotion	Ratio	0%
Years with current manager	Ratio	0%

III. Data Set Summary Statistics

Table 2 provides the statistical description of our variables used. In this description you will find the count, mean, minimum, 25th percentile, 50th percentile, 75th percentile, and the maximum. All of our numbers stayed the same from when we used python.

Table 2: Summary Statistics for XXX (name of dataset)

Variable Name	*Count*	*Mean*	*Min*	*25th*	*50th*	*75th*	*Max*

Age	*1470.00*	*36.92*	*18*	*30*	*36*	*43*	*60*
Distance From Home	*1470.00*	*9.193*	*1*	*2*	*7*	*14*	*29*
Employee Number	*1470.00*	*1024.87*	*1*	*491.25*	*1020.50*	*1555.75*	*2068.00*
Environment Satisfaction	*1470.00*	*2.72*	*1*	*2*	*3*	*4*	*4*
Hourly Rate	*1470.00*	*65.89*	*30.00*	*48.00*	*66.00*	*83.75*	*100.00*
Job Involvement	*1470.00*	*2.73*	*1*	*2*	*3*	*3*	*4*
Job Level	*1470.00*	*2.06*	*1*	*1*	*2*	*3*	*5*
Job Satisfaction	*1470.00*	*2.73*	*1*	*2*	*3*	*4*	*4*
Monthly Income	*1470.00*	*6502.93*	*1009.00*	*2911.00*	*4919*	*8379*	*19999*
Monthly Rate	*1470.00*	*14313.10*	*2094.00*	*8047*	*14235.5*	*20461.5*	*26999*
# Companies worked for	*1470.00*	*2.69*	*0*	*1*	*2*	*4*	*9*
Percent salary hike	*1470.00*	*15.21*	*11*	*12*	*14*	*18*	*25*
Performance Rating	*1470.00*	*3.15*	*3*	*3*	*3*	*3*	*4*
Total Working Years	*1470.00*	*11.28*	*0*	*6*	*10*	*15*	*40*
Training Times Last Year	*1470.00*	*2.80*	*0*	*2*	*3*	*3*	*6*
Years at Company	*1470.00*	*7.01*	*0*	*3*	*5*	*9*	*40*
Years in Current Role	*1470.00*	*4.23*	*0*	*2*	*3*	*7*	*18*
Years since last promotion	*1470.00*	*2.19*	*0*	*0*	*1*	*3*	*15*
Years With Current Manager	*1470.00*	*4.12*	*0*	*2*	*3*	*7*	*17*

Table 3: Proportions for Business Travel

Category	*Frequency*	*Proportion (%)*
Travel Rarely	*1043*	*70.9%*
Travel Frequently	*277*	*18.8%*
Non-Travel	*150*	*10.3%*

Proportions for Department

Category	*Frequency*	*Proportion (%)*
Research and Development	*961*	*65.4%*
Sales	*446*	*30.3%*
Human Resources	*63*	*4.3%*

Proportions for education field

Category	*Frequency*	*Proportion (%)*
Life Sciences	*606*	*41.2%*
Medical	*464*	*31.6%*
Marketing	*159*	*10.8%*
Technical Degree	*132*	*9.0%*
Other	*82*	*5.6%*
Human Resources	*27*	*1.8%*

Proportions for Gender

Category	*Frequency*	*Proportion (%)*
Male	*882*	*60%*
Female	*588*	*40%*

Proportions for Job Role

Category	*Frequency*	*Proportion (%)*
Sales Executive	*326*	*22.2%*
Research Scientist	*292*	*20.0%*
Laboratory Technician	*259*	*17.6%*
Manufacturing Director	*145*	*10.0%*
Healthcare Representative	*131*	*8.9%*
Manager	*102*	*6.9%*
Sales Representative	*83*	*5.6%*
Research Director	*80*	*5.4%*
Human Resources	*52*	*3.5%*

Proportions for Marital Status

Category	*Frequency*	*Proportion (%)*
Married	*673*	*45.8%*
Single	*470*	*32.0%*
Divorced	*327*	*22.2%*

Proportions for Over 18

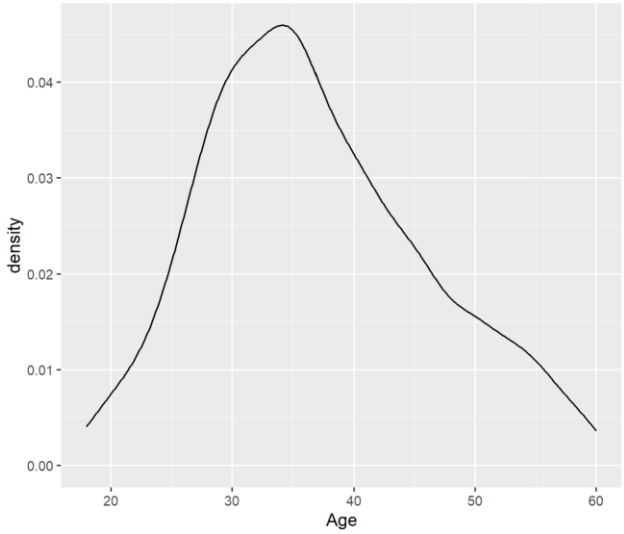
Category	*Frequency*	*Proportion (%)*
Yes	*1470*	*100%*

Proportions for Overtime

Category	*Frequency*	*Proportion (%)*
No	*1054*	*71.7%*
Yes	*416*	*28.3%*

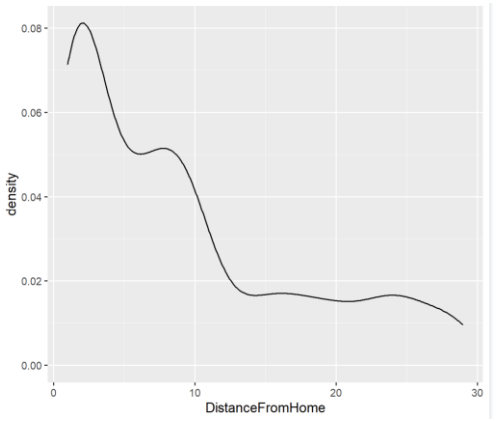
Our Plots:

Plot 1:



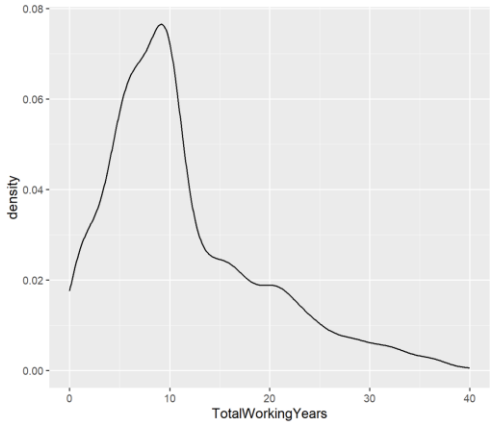
This is our density plot for age at this company. As shown above the age distribution at this company is skewed right meaning this company has more employees under the age of 40 than above 40.

Plot 2:



This is our density plot for "Distance from Home." This density plot shows a right skew with the majority of employees living very close to work. It is also close to bimodal but not fully there.

Plot 3:



This is our last density plot showing the density of "Total Working Years." This plot is also right skewed with most employees having worked under 15 years.

IV. DATA SET EXPLORATION

This section shows our comparisons of Age vs Hourly Rate, Total Working Years vs Hourly Rate, and Distance from Home vs Hourly Rate. We will go in depth for each investigation we did including the mean square error, simple linear regression formula, r-squared value, and a plot for each.

Investigation 1 (Age vs Hourly Rate):

During this investigation we found our values to be:

R-Sqaured – .06% This means that this is not a good prediction model.

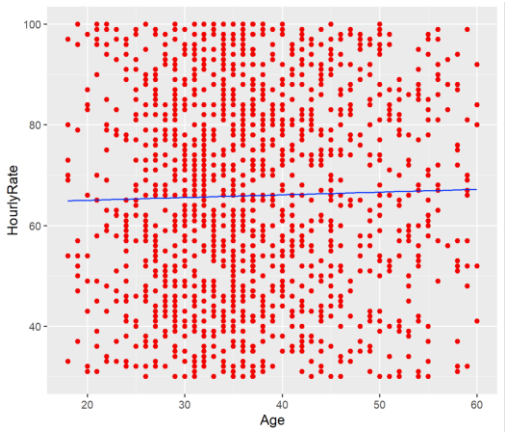
Linear Regression Formula – 63.896 + .0541*Age

Mean Squared Error – 412.76

P-Value – 0.3251 This also shows that we have a poor model to predict Hourly Rate using Age.

Prediction (Age 45) – Our model predicts that an individual at the age of 45 will have an hourly rate of $66.33.

Plot:



Investigation 2 (Total Working Years vs Hourly Rate):
In this investigation our values were:

R-Sqaured – 5.446e-06, This value means that we have a bad model for predicting Hourly Rate.

Linear Regression Formula – 65.9599 - .0061 * TotalWorkingYears

Mean Squared Error – 413.00

P-Value – 0.9288 This also shows that we have a very poor model to predict Hourly Rate using Total Working years.

Prediction (9 years) – Our model predicts that an employee with 9 years of experience will have an hourly rate of $65.91.

Plot:

Investigation 3 (Distance from Home vs Hourly Rate):
Our values for this investigation are:

R-Sqaured – .097% This means that this is not a good prediction model.

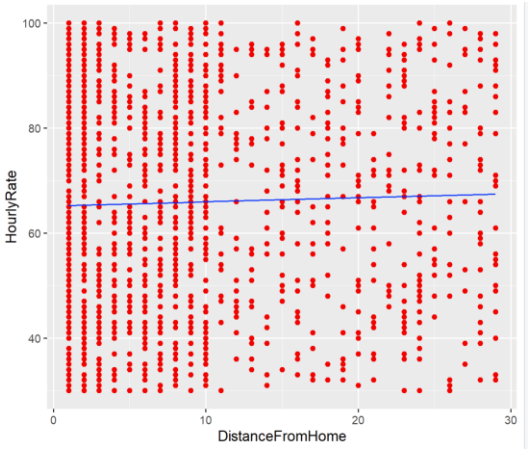
Linear Regression Formula – 65.1735 + .078*DistanceFromHome

Mean Squared Error – 412.60

P-Value – 0.2329 This also shows that we have a poor model to predict Hourly Rate using Distance from Home.

Prediction (15 miles) – Our model predicts than an employee with a 15-mile commute would have an hourly rate of $66.34.

Plot:



V. SUMMARY OF OUR FINDINGS

Overall, none of our investigations provided a strong model to predict Hourly Rate. The main reason for this is that there were many different types of jobs in our dataset. We also had skewed distributions in all of our density plots making it harder to make a good prediction. If we were to do our investigations again we would separate the job types into separate data categories which might help improve our prediction models.