## Introduction

This report presents the findings of a publicly available HR dataset that contains 311 observations of 39 variables. Rather than a business enabler, the human resources department is often looked at as a cost center; through the business questions posed below, I aimed to look at solutions that would reduce overall labor expenses in the long run, rather than the short run. After reviewing and cleaning the data, summary statistics and plots were formed as the basis for support vector machine and linear regression modeling. Some important measures in this dataset are gender, salary, turnover, employee satisfaction, and special project counts. All active employees in this dataset were full-time and an assumption was made that terminated employees were full-time as well. This dataset was last updated October 2020.

## Business Questions:

1. Where is turnover occurring and what is its impact?
2. Where are absences occurring and what is its impact?
3. Is there evidence of discrimination or mistreatment?
4. What strategies can be put into place to reduce or defer termination and absenteeism?

## Data Cleanse and Preparation

Methodology: after reading in and looking at the dataset, there were only N/A’s in the fields where it was appropriate (no termination date for active employees, no termination reason for active employees); however, a number of variables would have to be created in order to answer my business questions. Those are as follows:

1. Years of service would be derived from their termination date or a standard 2020-10-31 date (the last time the data was updated) if there was no termination date as they would still be active until otherwise listed. To create this variable, I had to re-format the Date of Hire and Date of Termination variables to be Y-M-D and then run them through the if\_else() function combined with a difftime() function that either used the 2020-10-31 date or termination date and stored the output in the Years of Service variable. I then had to remove the word “days” from the out, divide by 365, and then convert that to numeric.
2. Years Since Last Performance Review was created using the same reformat with a difftime() methodology above.
3. Age was created by that format and difftime() methodology as well.
4. Cost of Absence was derived from dividing the salary by 2080 to get an hourly rate, multiplied by 8 to get a daily rate, and then multiplied by the number of absences to get the employees total cost of their absences.
5. GenderID, MarriedID, and Termd were changed to factors using as.factor()
6. PerformanceID was not accurately reflecting the Performance Score string so it was mutated using case\_when() to be corrected.

The final variables and their structure are:

A screenshot of a computer

Description automatically generated

## Summary Statistics:

* summary(hrData)
  + 311 observations of 41 variables
  + 207 active employees, 104 terminated
  + 3184 absences costing $859,699.70
  + Mean salary: $69,021
  + Mean Engagement Survey Response: 4.11 out of 5
  + Mean Employee Satisfaction Response: 3.89 out of 5
  + Mean count of special projects: 1.219
    - Median 0
    - Max 8
  + Mean Absences: 10.24
  + Mean years of service: 8 years
  + Mean years since last performance review: 6.27 years
  + Mean Age: 44.89
  + Mean cost of Absence: 27624.3
    - Max: $13,566.2
  + 6 departments

A screenshot of a computer

Description automatically generated

### Terminated summary statistics:

Understanding where turnover is occurring.

*0 = active, 1=termd*

* + Terminations by department

A screenshot of a computer code

Description automatically generated

* + Termination Status by Gender

A screenshot of a computer code

Description automatically generated

### Absence Summary Statistics

* Sum of Absences by Department

A screenshot of a computer

Description automatically generated

* Sum of Absences by department>sex>termination status

A screenshot of a computer

Description automatically generated

From the summary data, I can see the production department has the largest amount of turnover and absences, being both 4x larger in its represented employees, 8x higher turnover, and 4x absences over the next largest department, IT. Some additional findings are:

* A proportionate total amount of active and terminated male and female values
* Active employees in production account for more absences than those terminated.

## Plots

To begin answering the second business question of where and how much absences are costing the business, I plotted the sum of the cost of absences by department

A graph with different colored squares

Description automatically generated

This revealed that Production had the largest cost, at roughly $500k, following IT/IS at $200k. I took a deeper dive and broke down the Production department by Position

A screenshot of a computer

Description automatically generated

This revealed that 60% of the departments cost, and 35% of the total business cost of absences, could be attributed to Production Technician 1. As I aimed to identify any gender discrimination, I looked at the above plot by gender

A graph of a graph

Description automatically generated with medium confidence

Which reveals roughly 60% of the cost of that position was attributed to female Production Technician I’s.

I also looked at where absences were occurring in the employee lifecycle by years of service which revealed:

A graph of different colored bars

Description automatically generatedAbsences are a problem of senior employees, which is not a-typical. But if we look at the same plot by termination status, we see:

A graph of a number of people

Description automatically generated with medium confidence

Employees stopped leaving around the 5-7 year mark and absences spiked. One possible interpretation of this is a personnel/staffing issue that has likely occurred for a few years, causing burnout and absenteeism among senior staff. The fact that absences in the above chart are occurring on non-terminated works could also indicate the inability to enforce attendance policy or an ineffective policy.

Looking at absences by termination status, I broke down the termination status into employment status, which broke down the 1-0 into active, terminated for cause, or voluntarily terminated. The count plot for this revealed:

A graph of a graph

Description automatically generated with medium confidence

Absences are not resulting in for cause terminations; rather, the absences of terminated employees are those that voluntarily left the company. This supports the prior notion that enforcement may not be occuring or policy strict enough to prevent individuals from simply being absent.

Having depth knowledge of the HR field, absences can occur due to employee satisfaction. The count plot for absences based on employee satisfaction revealed:

A graph of black dots

Description automatically generated

The majority of employees are average or above satisfied regardless of their absenteeism.

When satisfaction is not a factor, I would then look at their martial and number of dependent status. Employees may be satisfied with their work but family obligations may make them more absent. Dependent information was not available in this dataset but the boxplot of marital status count plot revealed:

A graph of dots on a white background

Description automatically generated

Married and single people are the largest populations represented and have consistent representation across the y-axis; not necessarily indicating a potential issue with one entity, but could represent a flaw, or area of improvement in the PTO/time off benefit offerings.

Too see if there were any gender trends to absenteeism, I looked at absences by gender by Manager, although Manager Gender was not an available field, looking at the names of managers, the column chart of this plot reveals:

A graph of a bar chart

Description automatically generated with medium confidence

In managers with assumed male names, there is often a larger portion of females absent; the same with assumed female manager names, males were more absent. Due to representation in this dataset, it is not a clear indicator of gender discrimination; however, it is worth noting as an opportunity for managerial training and understanding that of the opposite sex.

As a side; I looked at binned salary and mean absences which revealed:

A graph of a graph of a line

Description automatically generated with medium confidence

As salary increased, absences tended to increase at an almost equal slope between males and females. Although females are more compensated than males, pay does not appear to be a driving factor in gender absenteeism.

## Support Vector Machine Modeling: Turnover

To further analyze terminations, a support vector machine model was created with the variables: salary, engagement survey score, employee satisfaction, count of special projects, years since last performance review, days tardy in the last 30 days, and absences to predict whether an employee would terminate or not. These variables were chosen based on a business’ ability to control them, as well as events in the normal course of an employment cycle. ‘svmLinear’ was the best performing method for this model was 96.77% accurate, a roughly 23% increase over the no information rate.

A screenshot of a computer

Description automatically generated

Looking into the variable importance measure for the model:

A screenshot of a computer

Description automatically generated

Years since last performance review was by the most important variable in the model. To keep this leveled, all variables except YearsSinceLastPerfRev and Salary were removed. Performance was increased to 98.39%. Recommendation would be to identify the 2 incorrect predictions where they are not termed but predicted to be to assess how retention could occur.

## Linear Regression Model: Maximizing Years of Service

To find out what the important factors are in order maximize years of service, several linear regression models were run.

### Model 1

Was run with the independent variables salary, engagement survey score, employee satisfaction score, count of special projects, years since last performance review, absences, date late last 30 days, gender, marital status, and performance score. The result of this regression was a p-value of 2.2e-16, an adjusted r-squared value of .583, and only 2 statistically significant independent variables: Count of Special Projects and Years since last performance review.

Model 2

Was run with the same variables as the SVM with the y variable changing from Termd to Years of service. The result of this regression was a p-value of 2.2e-16, an adjusted r-squared value of .5857, and still only 2 statistically significant variables: County of Special Projects and Years since last performance review.

Model 3

Was run using only the statistically significant variables Special Projects and Years since last performance review. The result of this model was a p-value of 2.2e-16 and an adjusted r-squared value of .5867.

Interpretation: from the models, I would recommend using model 3 as the independent variables explained the largest amount of variation in the y-variable, years of service, at 58.67 but I would not recommend using a linear regression based on this set of data because while the model is statistically significant, it does not explain a lot of the outcome.

# Business Questions and Answers:

1&2. Where turnover and absences occurring and what is its impact?

* 1. There were 104 terminated employees in this dataset and production had the highest amount of termination at 83 people. Production also had the largest amount of absences with 2120 days and had a business cost of $493382.

These plots are the best summary of the impact of turnover and absenteeism is having.

A graph of a number of people

Description automatically generated with medium confidence

A graph of different colored bars

Description automatically generated

In the top plot, we see all turnover and absences occurring in the first 3-4 years of an employees service tenure and then at the 5-6 year mark, we see absences spike but little to no terminations. In this instance, we are seeing such high turnover early on in an employment lifecycle that absenteeism policy probably cannot be forced due to a lack of support staff to continue the workload. This has likely compounded for a couple of years.

1. Is there evidence of gender-based discrimination or mistreatment?
   1. It is hard to say if there is discrimination or mistreatment, but there appears to be a relationship between manager’s assumed gender and the gender with the most absences.
2. What strategies can be put into place to reduce or defer termination?
   1. To address the turnover situation, it is my recommendation that management adjusts its employment lifecycle strategy to include more frequent assignments of special projects and annual reviews. When employees had special projects, they had far less absences than if those that did not.

A graph with a bar graph

Description automatically generated

Roughly 20% of employees had special projects. When looking at special project assignments by department, we see an overwhelming bias to IT/IS.

A graph of a graph

Description automatically generated

To illustrate the impact special projects could have on reducing a departments turnover, IT/IS is 1/4th the size of Production and has 1/8th of the turnover.

In both the SVM and regression models, special projects and performance reviews variables were shown to be the one of the more statistically significant onces in affecting turnover and length of service which confirms recent workplace trends of employees wanting individualized assignments and more frequent feedback.

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generated

A number and numbers on a white background

Description automatically generated

By increasing the frequency in which performance reviews occur, managers and employees will be able to increase the dialogue on career development, progression, and interests which, when coupled with an increase in special projects, should strongly reduce or delay turnover and absenteeism within an employee’s first 3 years of the employees service length. In terms of a business impact, we see the coefficient for years since last performance review as -.89797 which is interpreted as for every year gone without a performance review, years of service is decreased by .89797 years. In the dataset, the median value is 1.78 years; this illustrates that a large number of employees had separated without having a review, or have much time to integrate the feedback (if any) from the review to their work schedule.

Special Projects having a negative coefficient is interesting; there were not that many individuals assigned special projects, the negative coefficient could suggest that people used those as career stepping stones outside of the organization. Turnover within an organization will always happen; if special projects were to become normalized, I would expect that number to be reduced or even become a positive value increasing the years of service length.

* 1. To address the absences and any evidence of mistreatment, I would recommend tracking absence reasons. This would help clarify what and where the issues are occurring. For example, if the absences were due to the care of a dependent, perhaps more remote or hybrid friendly work options would decrease the number of absences; lowering the cost of missed work. It could also reveal manager dissatisfaction. Not knowing what questions were asked in the employee satisfaction survey, there is not much to draw from as to why employees were satisfied but not present but additional data and strategy on the issues would supplement our understanding of the current status.

## Code:

library(tidyverse)

library(dplyr)

library(readr)

library(stringr)

library(caret)

library(kernlab)

library(rpart.plot)

library(ggplot2)

library(e1071)

hrData <-read\_csv(choose.files())

##Dataset and data dictionary can be found:

#https://www.kaggle.com/datasets/rhuebner/human-resources-data-set?resource=download

summary(hrData)

##Creation of Variables----------------------------------------------------------------------------------------------------

##creation of Years of Service requires reformatting date of hire to that of date of termination, using the difftime fuction to get

#days between 2 dates, stripping the days out of the field, converting to numeric format from char, and then dividing by 365

hrData$DateofHire <- format(as.Date(hrData$DateofHire, format = "%m/%d/%Y"), "%Y-%m-%d")

hrData$DateofTermination <- format(as.Date(hrData$DateofTermination, format = "%m/%d/%Y"), "%Y-%m-%d")

##2020-10-31 chosen as that was the last dataset update.

hrData$YearsofService <- if\_else(

hrData$Termd == 0,

difftime('2020-10-31',hrData$DateofHire, units = "days"),

difftime(hrData$DateofTermination, hrData$DateofHire, units = "days"))

hrData$YearsofService <- as.numeric(str\_replace(hrData$YearsofService, "days",""))/365

view(hrData)

##Creation of Days since last Perf Review

hrData$LastPerformanceReview\_Date <- format(as.Date(hrData$LastPerformanceReview\_Date, format = "%m/%d/%Y"), "%Y-%m-%d")

hrData$YearsSinceLastPerfRev <- if\_else(

hrData$Termd == 0,

difftime(Sys.Date(),hrData$DateofHire, units = "days"),

difftime(hrData$DateofTermination, hrData$LastPerformanceReview\_Date, units = "days"))

hrData$YearsSinceLastPerfRev<- difftime('2020-10-31', hrData$LastPerformanceReview\_Date, units = "days")

hrData$YearsSinceLastPerfRev <- as.numeric(str\_replace(hrData$YearsSinceLastPerfRev, "days",""))/365

##creation of age

hrData$DOB <- format(as.Date(hrData$DOB, format = "%m/%d/%Y"), "%Y-%m-%d")

hrData$DOB <- as.Date(hrData$DOB)

hrData$Age <- as.numeric(str\_replace(difftime(Sys.Date(), hrData$DOB, units = "days"),"days",""))/365

hrData$CostofAbsence <- (hrData$Salary/2080)\*8\*hrData$Absences

hrData$GenderID <- as.factor(hrData$GenderID)

hrData$MarriedID <- as.factor(hrData$MarriedID)

hrData<- hrData%>% mutate(PerfScoreID = case\_when(

PerformanceScore =='PIP'~'1',

PerformanceScore =='Needs Improvement'~'2',

PerformanceScore =='Fully Meets'~'3',

PerformanceScore =='Exceeds'~'4'

)

)

hrData$PerfScoreID <- as.numeric(hrData$PerfScoreID)

view(hrData)

str(hrData)

##Summary Information--------------------------------------------------------------------------------------------------------

summary(hrData)

sum(hrData$CostofAbsence)

hrData%>%group\_by(Termd)%>%summarise(n())

hrData%>% group\_by(Department)%>% summarise(n())

hrData%>%group\_by(Department)%>%summarise(sum(Termd))

hrData%>%group\_by(Sex, Termd)%>%summarise(n())

hrData%>%summarise(sum(Absences))

hrData%>%group\_by(Department)%>%summarise(sum(CostofAbsence))

hrData%>%group\_by(Department)%>%summarise(mean(EmpSatisfaction))

abs\_dep\_sex\_term<- hrData%>%group\_by(Department, Sex, Termd )%>%summarise(Sum\_abs = sum(Absences))

abs\_dep\_sex\_term <- abs\_dep\_sex\_term[order(abs\_dep\_sex\_term$Sum\_abs, decreasing = TRUE),]

abs\_dep\_sex\_term

mean(hrData$YearsofService)

summary(hrData$YearsSinceLastPerfRev)

##ggplots---------------------------------------------------------------------------------------------------------------------

hrData%>%filter(Termd == 1)%>%group\_by(YOS = YearsofService, Termd = as.factor(Termd))%>%summarise(n())%>% ggplot(aes(YOS)) + geom\_histogram()+labs(title = 'Termination Counts by YOS')

hrData%>%filter(TermReason !='N/A-StillEmployed')%>% ggplot(aes(TermReason)) +geom\_bar() +theme(axis.text.x=element\_text(angle=90, hjust=1))

view(hrData %>% group\_by(Position, Termd) %>% summarise(n(),sum(Absences), mean(Absences), mean(YearsofService)))

#col chart of cost of absences by department

ggplot(hrData, aes(x=Department, y=CostofAbsence,fill = Department, label = CostofAbsence)) + geom\_col() + scale\_y\_continuous(labels = scales::dollar\_format()) +labs(title = 'Cost of Absence by Department')

#col chart of cost of absences by position within the production department

hrDataProduction <- filter(hrData, hrData$Department == "Production")

ggplot(hrDataProduction) + geom\_col(aes(x = Position, y=CostofAbsence)) + coord\_flip()+scale\_y\_continuous(labels = scales::dollar\_format()) +labs(title = 'Cost of Absence by Production Department Position')

#col chart of cost of absences by position within the production department by gender

Plotvar1<- hrData %>% filter(Department == "Production") %>% group\_by(Position, Sex) %>% summarise(CostofAbsence = sum(CostofAbsence))

ggplot(Plotvar1, aes(x=Position, y = CostofAbsence, color = Sex, fill = Sex)) +geom\_col(position = 'stack') + coord\_flip()+scale\_y\_continuous(labels = scales::dollar\_format())+ labs(title = 'Cost of Position absences in the Production Department by Gender')

#count plot of employment status

ggplot(hrData) + geom\_count(aes(x = EmploymentStatus, y=Absences))

#boxplot of employee satisfaction absences

ggplot(hrData) + geom\_count(aes(x = as.factor(EmpSatisfaction), y = Absences))

#boxplot of marital status absences

ggplot(hrData) + geom\_count(aes(x = as.factor(MaritalDesc), y = Absences))

#are happier employees paid more? - Yes, they tend to be

ggplot(hrData) + geom\_boxplot(aes(x = as.factor(EmpSatisfaction), y=Salary))

view(hrData%>%group\_by(as.factor(GenderID))%>% summarise(IQR(Salary)))

hrData %>% group\_by(Absences = as.factor(Absences)) %>% summarise(AVG\_YOS = mean(YearsofService)) %>% ggplot(aes(Absences, AVG\_YOS)) + geom\_col()

#column chart of absences of gender of employees by manager name

ggplot(hrData, aes(x=ManagerName, y=Absences, fill=Sex)) + geom\_col() + coord\_flip()

#Mean absences by binned salary

ggplot(hrData,aes(x = Salary, y=Absences, color = Sex)) + geom\_point(stat="summary", fun = 'mean')+scale\_x\_binned(n.breaks = 25,labels = scales::dollar\_format())+theme(axis.text.x=element\_text(angle=90, hjust=1)) +geom\_smooth(method = 'lm')

#bar chart count of employee sex

ggplot(hrData, aes(as.factor(x=Sex))) + geom\_bar()

#col chart of absences by sex

ggplot(hrData) + geom\_col(aes(x=Sex, y=Absences)) + labs(title = 'Absences by Sex')

#looking at avg years of service to see if it was a recruiting issue - it wasnt

hrData %>% group\_by(RecruitmentSource) %>% summarise(AVG\_YOS = mean(YearsofService)) %>% ggplot(aes(RecruitmentSource, AVG\_YOS)) +

geom\_col()

#Looking to see if there were any departments with lower years of service avergaes - there werent

hrData %>% group\_by(Department) %>% summarise(AVG\_YOS = mean(YearsofService)) %>% ggplot(aes(Department, AVG\_YOS)) +geom\_col()

view(hrData%>%group\_by(as.factor(GenderID))%>% summarise(IQR(Salary)))

ggplot(hrData, aes(x=YearsofService, y=Absences, fill = Department)) + geom\_col() + scale\_x\_binned(n.breaks = 15)

ggplot(hrData, aes(x=YearsofService, y=Absences, fill = as.factor(Termd))) + geom\_col() + scale\_x\_binned(n.breaks = 15)

ggplot(hrData, aes(x=YearsofService, y=Absences, fill = Sex)) + geom\_col() + scale\_x\_binned(n.breaks = 15)

ggplot(hrData, aes(x=Department, y=SpecialProjectsCount))+geom\_col()

hrData%>%group\_by(Department)%>% summarise(sum(SpecialProjectsCount), sum(Absences))

##SVM----------------------------------------------------------------------------------------------------------------------------

set.seed(400)

HRDataModel <- data.frame(

Termd = factor(hrData$Termd),

Salary = hrData$Salary,

# Engagement\_Survey = hrData$EngagementSurvey,

#Emp\_Satisfaction = hrData$EmpSatisfaction,

#Special\_Projects = hrData$SpecialProjectsCount,

YearsSinceLastPerfRev = hrData$YearsSinceLastPerfRev

#Absences = hrData$Absences,

#Tardy30 = hrData$DaysLateLast30,

)

trainTermd <- createDataPartition(y=hrData$Termd, p=.6, list = FALSE)

trainsetTermd <- HRDataModel[trainTermd,]

testsetTermd<- HRDataModel[-trainTermd,]

svm.model <- train(Termd~., data = trainsetTermd,

method = "svmLinear",

trcontrol=trainControl(method = 'none'),

preProcess=c("center","scale"))

svm.model$finalModel

svmPrediction <- predict(svm.model, testsetTermd, type = "raw")

x <- table(testsetTermd$Termd, svmPrediction)

confusionMatrix(x)

cbind(testsetTermd$Termd, svmPrediction)

varImp(svm.model)

svm.model$results

##Linear Regression--------------------------------------------------------------------------------------------------------------

HRDataModelLM1 <- data.frame(

Salary = hrData$Salary,

Engagement\_Survey = hrData$EngagementSurvey,

Emp\_Satisfaction = hrData$EmpSatisfaction,

Special\_Projects = hrData$SpecialProjectsCount,

YOS = hrData$YearsofService,

YearsSinceLastPerfRev = hrData$YearsSinceLastPerfRev,

Absences = hrData$Absences,

Tardy30 = hrData$DaysLateLast30,

Gender = as.factor(hrData$GenderID),

Marital\_Status = as.factor(hrData$MarriedID),

Performance\_score = hrData$PerfScoreID

)

lmYOS1 <- lm(YOS~., data = HRDataModelLM1)

summary(lmYOS1)

HRDataModelLM2 <- data.frame(

Salary = hrData$Salary,

Engagement\_Survey = hrData$EngagementSurvey,

Emp\_Satisfaction = hrData$EmpSatisfaction,

Special\_Projects = hrData$SpecialProjectsCount,

YearsSinceLastPerfRev = hrData$YearsSinceLastPerfRev,

Absences = hrData$Absences,

Tardy30 = hrData$DaysLateLast30,

YOS = hrData$YearsofService

)

lmYOS2 <- lm(YOS~., data = HRDataModelLM2)

summary(lmYOS2)

HRDataModelLM3 <- data.frame(

Special\_Projects = hrData$SpecialProjectsCount,

YearsSinceLastPerfRev = hrData$YearsSinceLastPerfRev,

YOS = hrData$YearsofService

)

lmYOS3 <- lm(YOS~., data = HRDataModelLM3)

summary(lmYOS3)