First MarkdOWn Document

2024-03-20

```
# install.packages("tinytex")
# tinytex::install_tinytex()
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
pay <- read.csv("pay_history_clean.csv")
head(pay)
```

```
##   EmployeeID DepartmentID RateChangeDate     Rate PayFrequency
## 1          1           16        00:00.0 125.5000            2
## 2          2            1        00:00.0  63.4615            2
## 3          3            1        00:00.0  43.2692            2
## 4          4            1        00:00.0   8.6200            2
## 5          4            2        00:00.0   8.6200            2
## 6          4            2        00:00.0  23.7200            2
##                     LoginID OrganizationLevel                     JobTitle
## 1     adventure-works\\ken0              NULL       Chief Executive Officer
## 2   adventure-works\\terri0                 1 Vice President of Engineering
## 3 adventure-works\\roberto0                 2           Engineering Manager
## 4     adventure-works\\rob0                 3          Senior Tool Designer
## 5     adventure-works\\rob0                 3          Senior Tool Designer
## 6     adventure-works\\rob0                 3          Senior Tool Designer
##    BirthDate MaritalStatus Gender   HireDate SalariedFlag VacationHours
## 1  1/29/1969             S      M  1/14/2009            1            99
## 2   8/1/1971             S      F  1/31/2008            1             1
## 3 11/12/1974             M      M 11/11/2007            1             2
## 4 12/23/1974             S      M  12/5/2007            0            48
## 5 12/23/1974             S      M  12/5/2007            0            48
## 6 12/23/1974             S      M  12/5/2007            0            48
##   SickLeaveHours CurrentFlag ShiftID  StartDate   EndDate ModifiedDate
## 1             69           1       1  1/14/2009      NULL      00:00.0
## 2             20           1       1  1/31/2008      NULL      00:00.0
```

```
## 3             21           1       1 11/11/2007      NULL      00:00.0
## 4             80           1       1  12/5/2007 5/30/2010      00:00.0
## 5             80           1       1  5/31/2010      NULL      00:00.0
## 6             80           1       1  5/31/2010      NULL      00:00.0
##   DepartmentName                       Sub.Department
## 1      Executive Executive General and Administration
## 2    Engineering             Research and Development
## 3    Engineering             Research and Development
## 4    Engineering             Research and Development
## 5    Tool Design             Research and Development
## 6    Tool Design             Research and Development
```

Summary Introduction:

This is payroll data or employment data from a company that hires people. The data contains a "Rate" of pay field, numeric data that I used. I had hoped to use the three dates provided, "BirthDate, StartDAte and EndDate," however, these columns proved to have "NULL" values that caused data loss. I have produced 8 graphic visualizations from this data.

```
summary(pay)
```

```
##    EmployeeID      DepartmentID    RateChangeDate          Rate
##  Min.   :  1.00   Min.   : 1.000   Length:304         Min.   :  6.50
##  1st Qu.: 72.75   1st Qu.: 7.000   Class :character   1st Qu.: 11.00
##  Median :148.50   Median : 7.000   Mode  :character   Median : 14.00
##  Mean   :146.83   Mean   : 7.303                      Mean   : 18.15
##  3rd Qu.:224.00   3rd Qu.: 7.000                      3rd Qu.: 23.08
##  Max.   :290.00   Max.   :16.000                      Max.   :125.50
##   PayFrequency     LoginID          OrganizationLevel    JobTitle
##  Min.   :1.000   Length:304         Length:304         Length:304
##  1st Qu.:1.000   Class :character   Class :character   Class :character
##  Median :1.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1.461
##  3rd Qu.:2.000
##  Max.   :2.000
##   BirthDate         MaritalStatus         Gender            HireDate
##  Length:304         Length:304         Length:304         Length:304
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   SalariedFlag    VacationHours   SickLeaveHours   CurrentFlag    ShiftID
##  Min.   :0.0000   Min.   : 0.00   Min.   :20.00   Min.   :1    Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:26.75   1st Qu.:33.00   1st Qu.:1    1st Qu.:1.000
##  Median :0.0000   Median :49.00   Median :45.00   Median :1    Median :1.000
##  Mean   :0.2039   Mean   :49.96   Mean   :45.21   Mean   :1    Mean   :1.546
##  3rd Qu.:0.0000   3rd Qu.:74.00   3rd Qu.:57.25   3rd Qu.:1    3rd Qu.:2.000
##  Max.   :1.0000   Max.   :99.00   Max.   :80.00   Max.   :1    Max.   :3.000
##   StartDate           EndDate          ModifiedDate       DepartmentName
##  Length:304         Length:304         Length:304         Length:304
##  Class :character   Class :character   Class :character   Class :character
```

```
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Sub.Department
##  Length:304
##  Class :character
##  Mode  :character
##
##
##
```

Including Plots

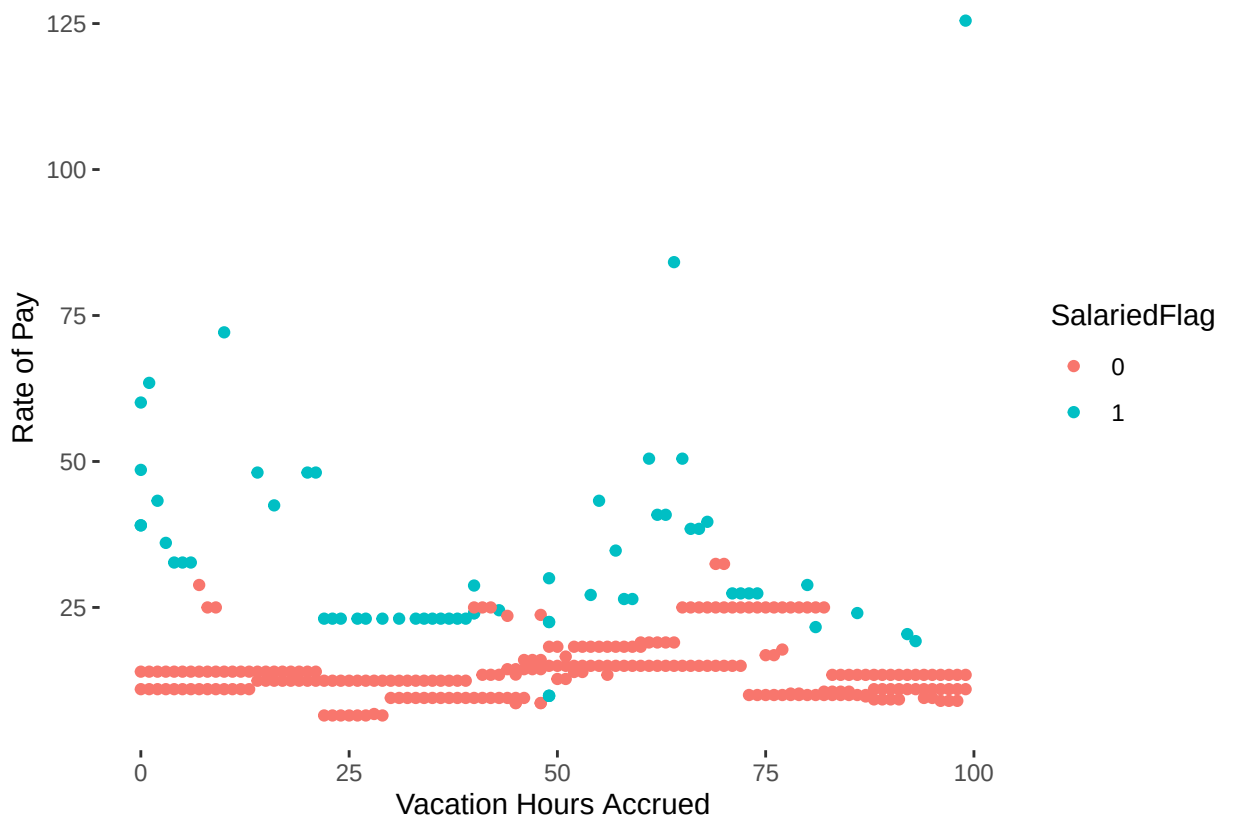
```
pay <- pay %>%
  mutate(EmployeeID=as.character(EmployeeID), SalariedFlag=as.character(SalariedFlag))
str(pay)
```

```
## 'data.frame':    304 obs. of  22 variables:
##  $ EmployeeID       : chr  "1" "2" "3" "4" ...
##  $ DepartmentID     : int  16 1 1 1 2 2 1 1 6 6 ...
##  $ RateChangeDate   : chr  "00:00.0" "00:00.0" "00:00.0" "00:00.0" ...
##  $ Rate             : num  125.5 63.46 43.27 8.62 8.62 ...
##  $ PayFrequency     : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ LoginID          : chr  "adventure-works\\ken0" "adventure-works\\terri0" "adventure-works\\rober
##  $ OrganizationLevel: chr  "NULL" "1" "2" "3" ...
##  $ JobTitle         : chr  "Chief Executive Officer" "Vice President of Engineering" "Engineering Ma
##  $ BirthDate        : chr  "1/29/1969" "8/1/1971" "11/12/1974" "12/23/1974" ...
##  $ MaritalStatus    : chr  "S" "S" "M" "S" ...
##  $ Gender           : chr  "M" "F" "M" "M" ...
##  $ HireDate         : chr  "1/14/2009" "1/31/2008" "11/11/2007" "12/5/2007" ...
##  $ SalariedFlag     : chr  "1" "1" "1" "0" ...
##  $ VacationHours    : int  99 1 2 48 48 48 5 6 61 62 ...
##  $ SickLeaveHours   : int  69 20 21 80 80 80 22 23 50 51 ...
##  $ CurrentFlag      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ ShiftID          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ StartDate        : chr  "1/14/2009" "1/31/2008" "11/11/2007" "12/5/2007" ...
##  $ EndDate          : chr  "NULL" "NULL" "NULL" "5/30/2010" ...
##  $ ModifiedDate     : chr  "00:00.0" "00:00.0" "00:00.0" "00:00.0" ...
##  $ DepartmentName   : chr  "Executive" "Engineering" "Engineering" "Engineering" ...
##  $ Sub.Department   : chr  "Executive General and Administration" "Research and Development" "Researc
```

Graph 1

This graph demonstrates that salaried employees earn a higher wage than non-salaried employees. Salaried employees may work additional hours that they are not necessarily paid by the hour for having worked them. Because this as an "accrued" bank of Vacation hours it is difficult to tell, who was hired most recently and therefore has fewer hours accrued? Who is taking vacations as opposed to those employees who don't take time off for years?

```
ggplot(data=pay) +
  geom_point(mapping = aes(x=VacationHours, y=Rate, color=SalariedFlag))  +     theme(panel.background =
  ylab("Rate of Pay") +
  xlab("Vacation Hours Accrued") +
  scale_fill_manual(guide=guide_legend(title="Employee Type"))
```

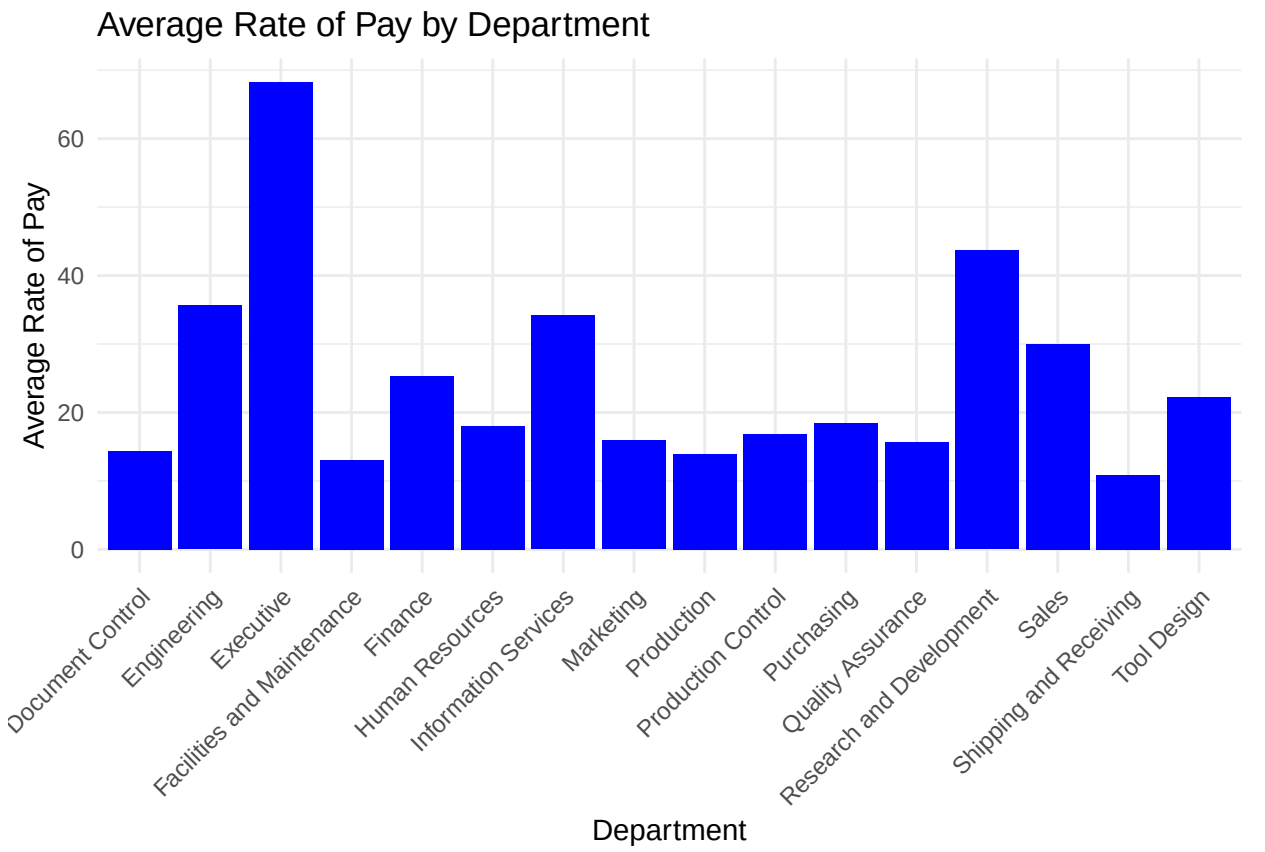


Graph 2

This bar graph groups average pay rates by category of job title. We see the tallest bar or greatest wage corresponds to executive level employees. Executives earn an average wage of $69 dollars per hour. This does not include any bonus earnings.Research and Development associates, on average $43/hour. The Engineering staff ranks third in wages at this company. I expected the Sales department to rank above $30/hr. I assume Sales receives bonus and commission that are not reflected in their baseline wage.

```
if(!requireNamespace("dplyr", quietly=TRUE)) install.packages("dplyr")
library(dplyr)
 avg_pay_rate_by_dept <-
  pay %>%
  group_by(DepartmentName) %>%
              summarise(avg_pay_rate_by_dept = mean(Rate, na.rm=TRUE))

  ggplot(avg_pay_rate_by_dept) +
    geom_col(mapping=aes(x=DepartmentName, y=avg_pay_rate_by_dept), fill="blue") +
    theme(panel.background = element_blank(), legend.key = element_blank()) +
```

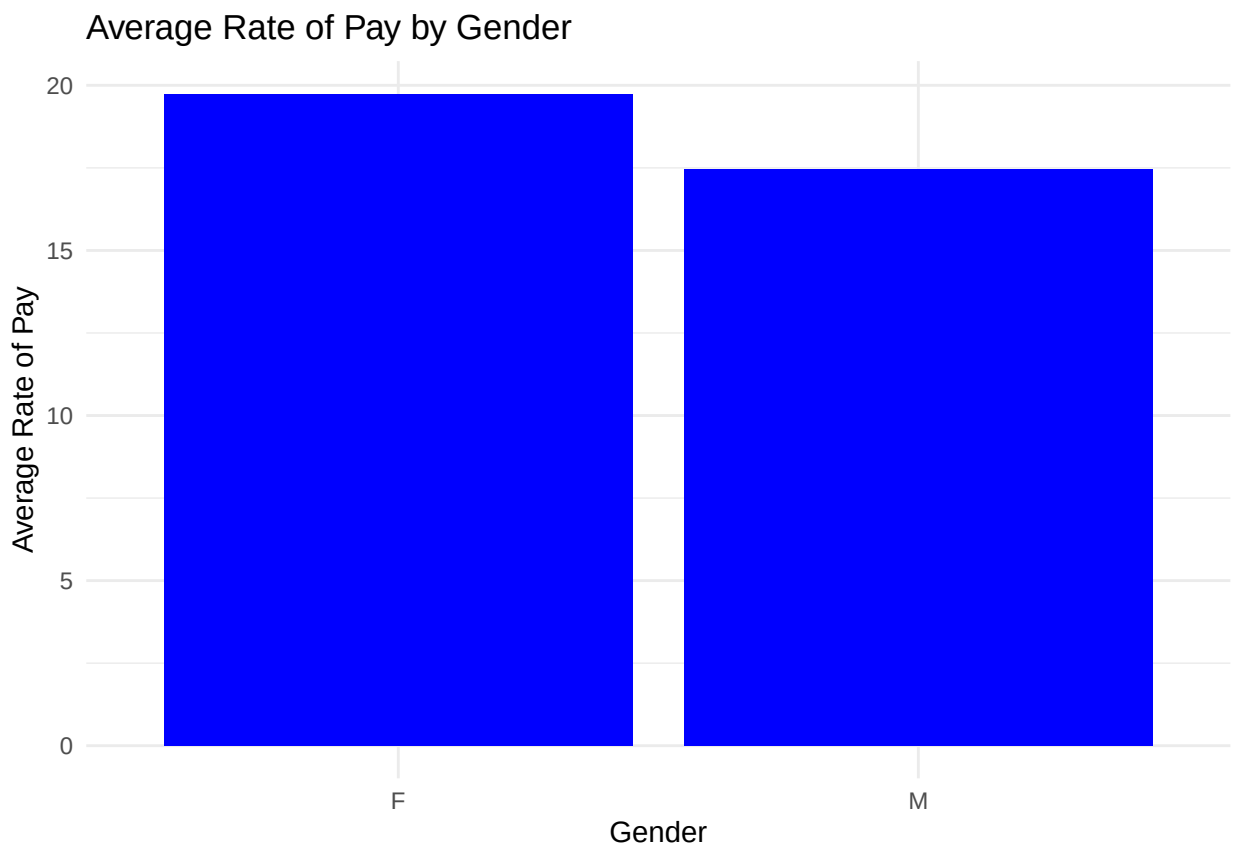
```
    theme_minimal() +
    labs(title = "Average Rate of Pay by Department",
         x="Department",
         y="Average Rate of Pay") +
    theme(axis.text.x=element_text(angle=45, hjust=1))
```



Graph 3

Females at this company earn about $2.50 more per hour on average than the males. Many other factors are at play here like the number of females that contribute to this average. Perhaps more women than men work for this company? The CEo of this company is male and they earn the highest salary of the whole business. I expected the male wage to be higher on average than the female wage.

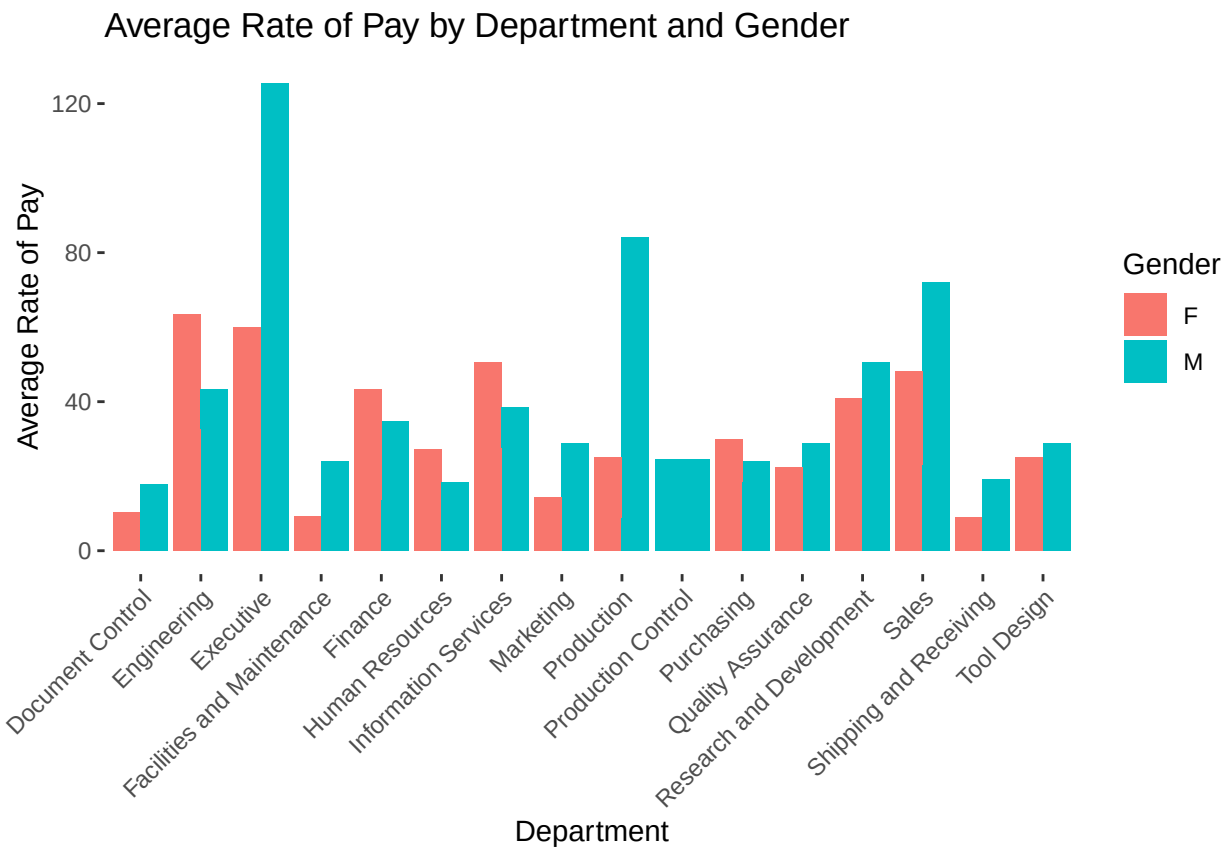
```
library(dplyr)
avg_rate_of_pay_by_gender <-
  pay %>%
  group_by(Gender) %>%
    summarize(avg_rate_of_pay_by_gender = mean(Rate, na.rm=TRUE))
ggplot(avg_rate_of_pay_by_gender) +
  geom_col(mapping=aes(x=Gender, y=avg_rate_of_pay_by_gender), fill="blue") +
  theme(panel.background = element_blank(), legend.key = element_blank()) +
  theme_minimal() +
    labs(title = "Average Rate of Pay by Gender",
         x="Gender",
         y="Average Rate of Pay")
```



Graph 4

This graph expands on the general nature of Graph 3 by separating out the female and male sex inside of each department. In the executive and production departments the male average rate of pay exceeds that of the females by over 50%! The male CEO is causing the executive male category to be the subject of an outlier. The departments that pay women a higher wage on average than the men, don't exceed the male average by more than $15 per hour.

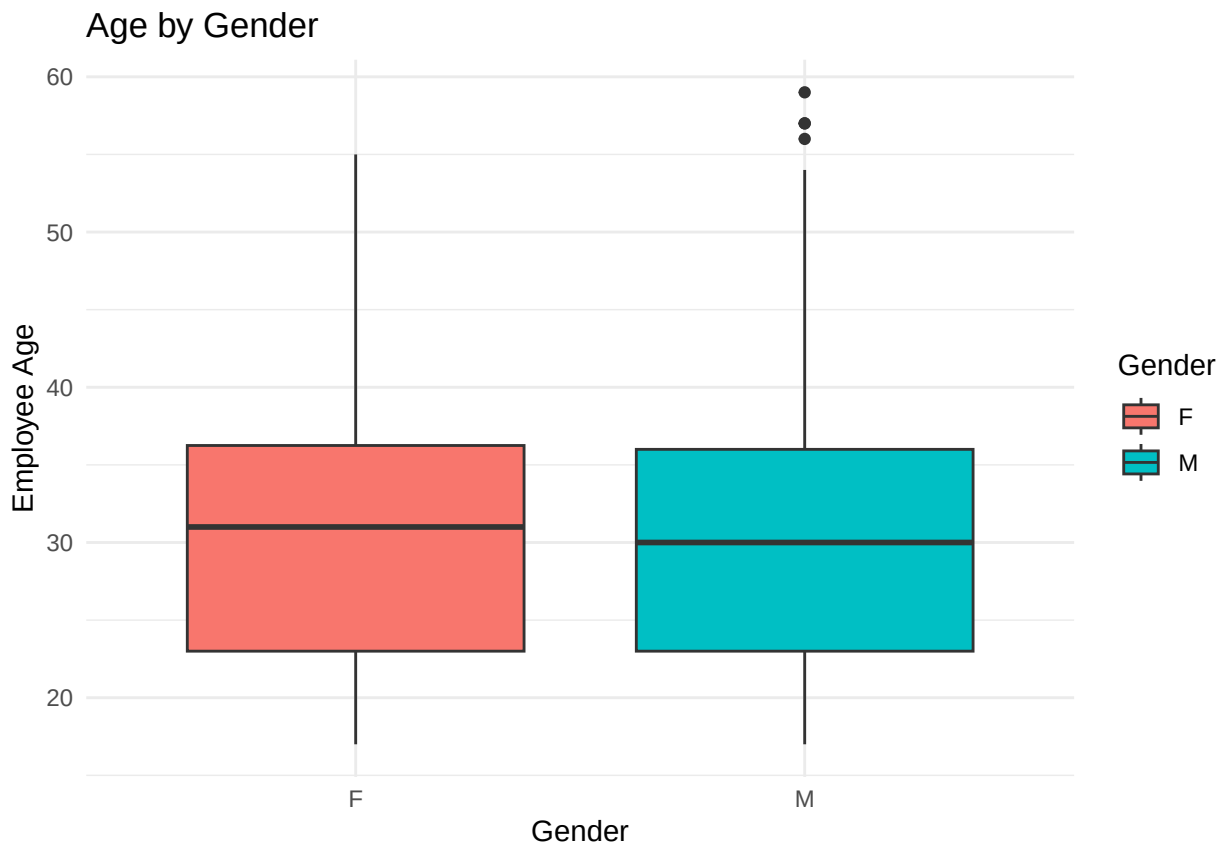
```
library(dplyr)
ggplot(pay, aes(x=DepartmentName, y=Rate, fill = Gender)) +
    geom_bar(stat = "identity", position = "dodge") +
    labs(title = "Average Rate of Pay by Department and Gender",
         x = "Department",
         y = "Average Rate of Pay",
         fill = "Gender") +
    theme(panel.background = element_blank(), legend.key = element_blank()) +
  theme(axis.text.x=element_text(angle=45, hjust=1))
```



Graph 5

I used a box and whisker to ascertain the relative age of the workforce at this company. This company employs a few males and females under the age of 20 years. This is a young workforce with the median age for females at 31 years and for males 30 years. The male group has three employees that age beyond the upper whisker indicating 3 outliers.

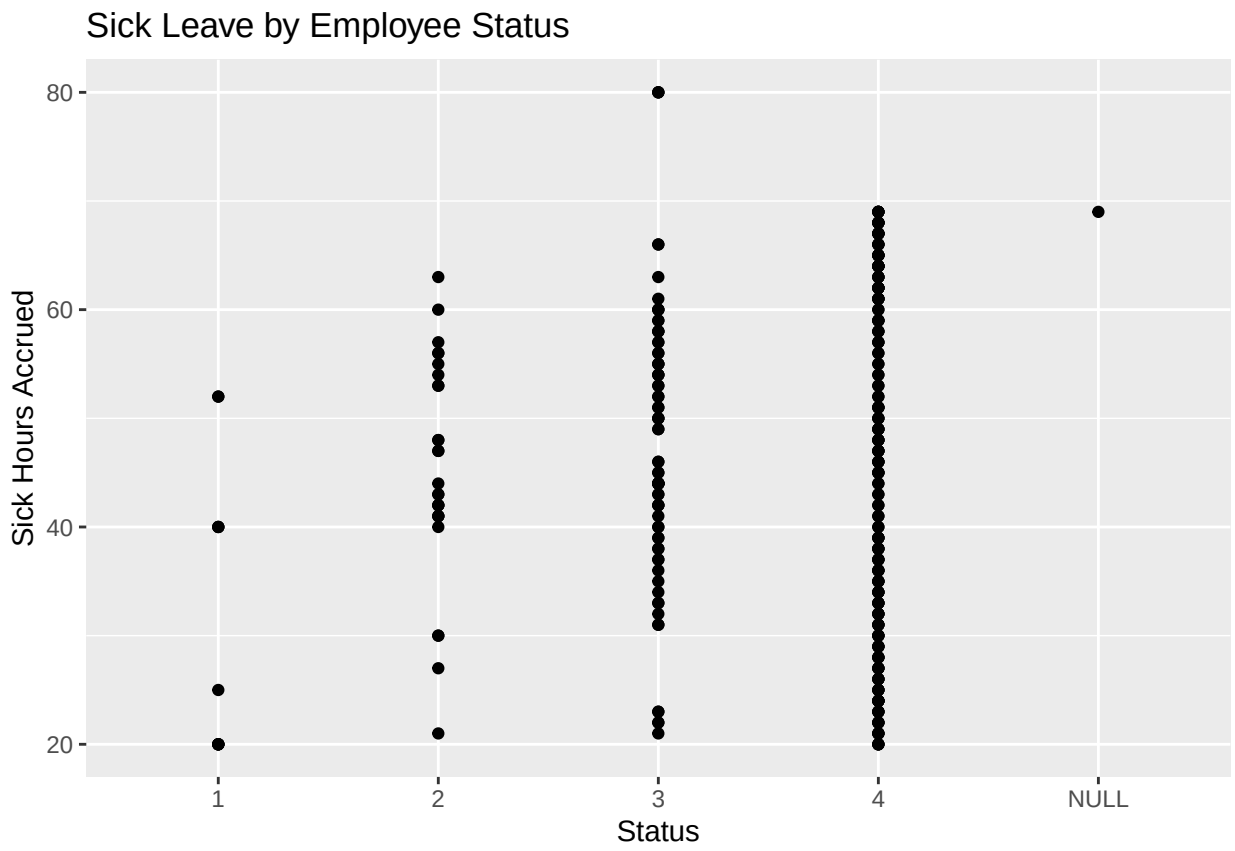
```
#  pay <- data.frame(
 pay$BirthDate <- as.Date(pay$BirthDate, "%m/%d/%Y")
 pay$StartDate <-  as.Date(pay$StartDate, "%m/%d/%Y")
 Gender <- c("Male", "Female")
 pay$age_of_employee <- as.integer(difftime(pay$StartDate, pay$BirthDate, units = "days") / 365.25)

library(ggplot2)
  ggplot(pay, aes(x=Gender, y=age_of_employee, fill=Gender)) +
    geom_boxplot() +
    labs(title="Age by Gender", x="Gender", y="Employee Age") +
    theme_minimal()
```



Graph 6

This data set lacked numerical data that could be used as a scatter to determine possible correlations. This graph "counts" the employee as categorized by their status to demonstrate their accrued sick leave hours. Assuming that Status 1 is a "lower" status worker than status 4, a couple of things could be happening here. The status 1 employee accrues less sick leave than a status 4 employee. The status 1 employee is of a lower socioeconomic status and is sick more frequently. Alternatively, they use their sick leave to care for family members who have no help to attend doctors appointments to obtain basic level care. Status 4 employees tend to get sick less because the graph shows more dots in level 4 than any other status level. I think the 1 NULL value is inconsequential to this analysis.

```
ggplot(pay, aes(x=OrganizationLevel, y=SickLeaveHours)) +
  geom_point() +
  labs(title="Sick Leave by Employee Status ", x="Status", y="Sick Hours Accrued")
```

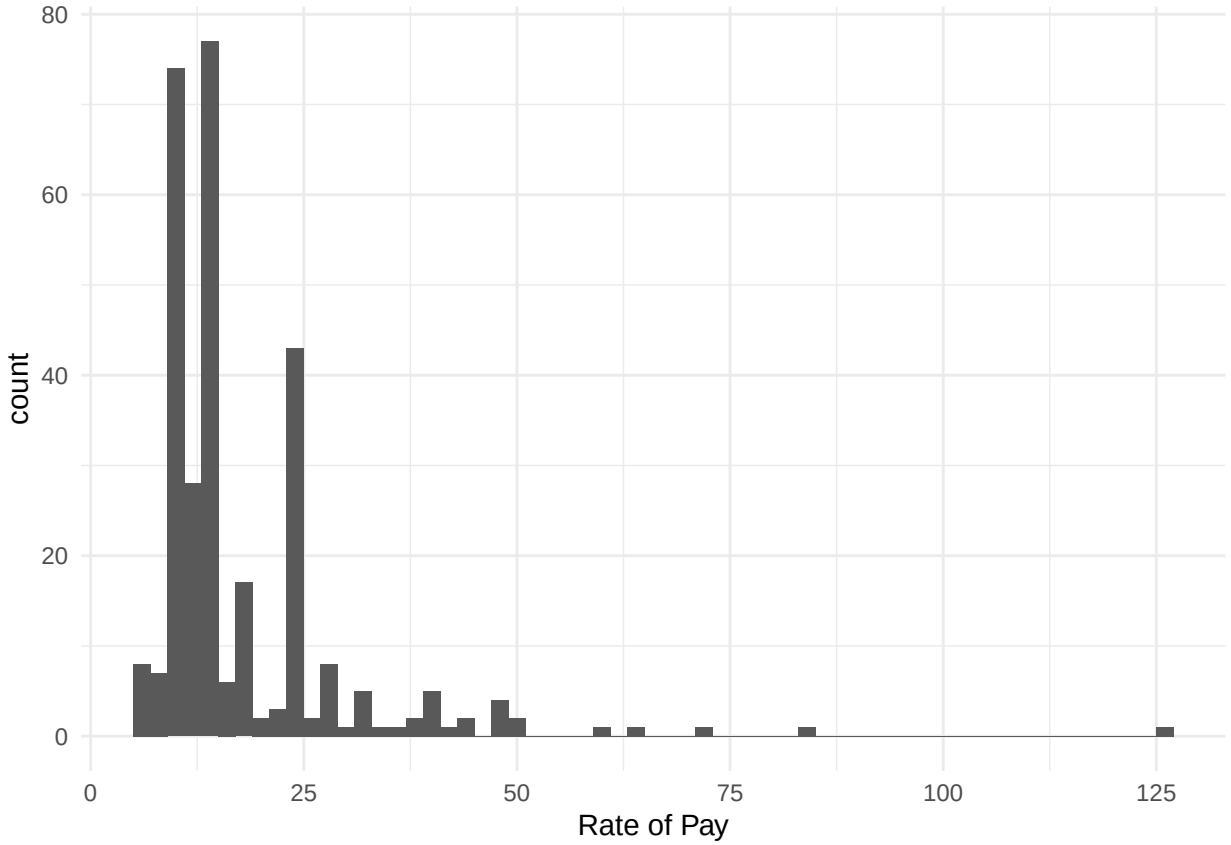



```
#ggplot(pay, aes(x=Gender, y=age_of_employee, fill=Gender)) +
 # geom_boxplot() +
  #labs(title="Age by Gender", x="Gender", y="Employee Age") +
  #theme_minimal()
```

Graph 7

This is a continuation of hourly pay rate on the x-axis into continuous bins using a histogram. The binwidth at 2 attempts to separate out the very large counts of wages that are very close together at the low end of wage. As the wage gets larger, few employees are in those wage categories as evidenced by the short bins. The axis markers are not super exact but $12.50 per hour is the most popular wage bin, plus or minus $2.

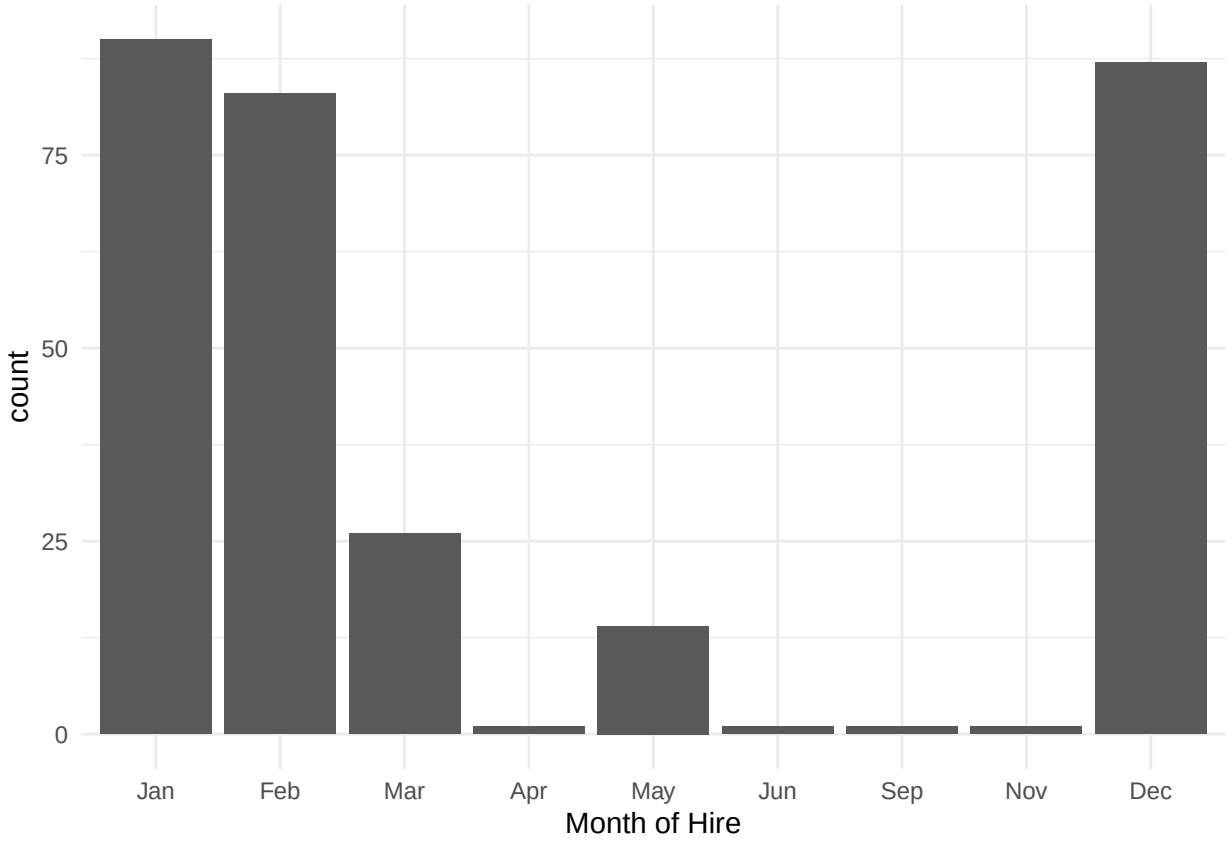
```
ggplot(pay, aes(x=Rate)) +
  geom_histogram(binwidth=2) +
  theme_minimal() +
  labs(x="Rate of Pay")
```



Graph 8

According to the HireDate of this companies employees, the majority of employees have been hired during the dead of Winter. If this was Christmas related I would expect to see hiring in October or November. More analysis would be needed to know if these are mostly production workers as this graph includes all employees. A seed, soil, or gardening type business might hire with this same pattern.

```
pay$HireMonth <- month(as.Date(pay$HireDate, "%m/%d/%Y"), label=TRUE, abbr=TRUE)
ggplot(pay, aes(x=HireMonth)) +
  geom_bar() +
  theme_minimal() +
  labs(x="Month of Hire")
```



Conclusion This business employs a young workforce and experiences a seasonal hiring surge during the December, January and February. They employ women and men in each department except for Production Control. Production Control is exclusively men. The majority of workers earn below $25 per hour. The highest status of worker has the most accrued sick leave hours suggesting the socio-economics afford them privilege. In the departments that male wages exceed female wages, that extra wages for men exceed the overages in departments when the females earn more than the men. For females the overage is marginal as opposed to 50% or greater. The vacation hours accrued are dispersed evenly across the average rate of pay making it difficult to know if accrual rates vary based on employee status.