Assignment 2

Code

Hide

```
myDataLocation <- "C:/Users/NoBull/Desktop/ANA600/R-Studio"
setwd(myDataLocation)
IBMdata <- read.csv(file = "IBMdata.csv", header = TRUE)

library(knitr)
library(ggplot2)
library(ggformula)
library(mosaic)
library(dplyr)
library(supernova)
library(car)
library(ltm)
library(lsr)
library(nycflights13)
library(fueleconomy)
library(palmerpenguins)
library(Lock5Data)
```

Part 1. Purpose In trying to decipher employee attrition rates (when an employee leaves the company), there are a few interesting hypotheses that can be explored:

1. Does education level affect employment length? Employees with a Doctorate, Master's, or Bachelor's degree have different average tenures compared to those with a High School diploma.

Years at Company = Education + other stuff

2. Does the number of years an employee stays at the company vary based on the time after their last promotion? Employees stuck in their position could feel job security.

Years at Company = Years Since Promotion + other stuff

3. Does business travel frequency impact work-life balance? Frequent business travel creates poor work-life balance for employees and cause them to leave.

Years at Company = Business travel + Work-Life balance + other stuff

```
'data.frame':   1470 obs. of  35 variables:
 $ ï..Age                  : int  41 49 37 33 27 32 59 30 38 36 ...
 $ Attrition               : chr  "Yes" "No" "Yes" "No" ...
 $ BusinessTravel          : chr  "Travel_Rarely" "Travel_Frequently" "Travel_Rarely" "Travel
_Frequently" ...
 $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
 $ Department              : chr  "Sales" "Research & Development" "Research & Development" "
Research & Development" ...
 $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
 $ Education               : int  2 1 2 4 1 2 3 1 3 3 ...
 $ EducationField          : chr  "Life Sciences" "Life Sciences" "Other" "Life Sciences" ...
 $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ EmployeeNumber          : int  1 2 4 5 7 8 10 11 12 13 ...
 $ EnvironmentSatisfaction : int  2 3 4 4 1 4 3 4 4 3 ...
 $ Gender                  : chr  "Female" "Male" "Male" "Female" ...
 $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
 $ JobInvolvement          : int  3 2 2 3 3 3 4 3 2 3 ...
 $ JobLevel                : int  2 2 1 1 1 1 1 1 3 2 ...
 $ JobRole                 : chr  "Sales Executive" "Research Scientist" "Laboratory Technici
an" "Research Scientist" ...
 $ JobSatisfaction         : int  4 2 3 3 2 4 1 3 3 3 ...
 $ MaritalStatus           : chr  "Single" "Married" "Single" "Married" ...
 $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
 $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 13335 8787 16577
...
 $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
 $ Over18                  : chr  "Y" "Y" "Y" "Y" ...
 $ OverTime                : chr  "Yes" "No" "Yes" "Yes" ...
 $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
 $ PerformanceRating       : int  3 4 3 3 3 3 4 4 4 3 ...
 $ RelationshipSatisfaction: int  1 4 2 3 4 3 1 2 2 2 ...
 $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
 $ StockOptionLevel        : int  0 1 0 0 1 0 3 1 0 2 ...
 $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
 $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
 $ WorkLifeBalance         : int  1 3 3 3 3 2 2 3 3 2 ...
 $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
 $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
 $ YearsSinceLastPromotion : int  0 1 0 3 2 3 0 0 1 7 ...
 $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
```

Part2: Data

a. **The following 10 variables are listed as categorical:**

Hide

```
list_character_variables <- function(dataframe) {
  char_vars <- names(dataframe)[sapply(dataframe, is.character)]
  return(char_vars)
}
character_variables <- list_character_variables(IBMdata)
print(character_variables)
```

```
 [1] "Attrition"       "BusinessTravel"  "Department"
 [4] "EducationField"  "Gender"          "JobRole"
 [7] "MaritalStatus"   "Over18"          "OverTime"
[10] "PromotionStatus"
```

There are an additional 11 variables listed as “int” that are in fact discreet categories:

“Education” “EmployeeCount” “EmployeeNumber”

“EnvironmentSatisfaction” “JobInvolvement” “JobLevel”

“JobSatisfaction” “NumCompaniesWorked” “PerformanceRating”

“RelationshipSatisfaction” “WorkLifeBalance”

The following 15 variables are quantitative:

“ï..Age” “DailyRate” “DistanceFromHome”

“HourlyRate” “MonthlyIncome” “MonthlyRate”

“PercentSalaryHike” “StandardHours” “StockOptionLevel”

“TotalWorkingYears” “TrainingTimesLastYear” “YearsAtCompany”

“YearsInCurrentRole” “YearsSinceLastPromotion” “YearsWithCurrManager”

b. The Years Since Last Promotion variable offers an opportunity to look at promotion rates and their effect on how long an employee stays.

c. All employees with 0 Since Last Promotion are listed as “New” to separate from “On track” employees. A secondary comparison with Years at company would clarify which category they should be put into. A look at promotion policy would also be needed in order to define the upper cut-off for “On Track”, but the large differentiation between level 2 and 3 would lend to a reasonable hypothesis that 1-2 are typical promotion periods. Employees above 3 years are considered passed over, but Job Role would probably need to be considered as well.

Hide

```
summary(IBMdata$YearsSinceLastPromotion)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   2.188   3.000  15.000
```

Hide

```
IBMdata <- IBMdata |>
    mutate(
      PromotionStatus = case_when(
      YearsSinceLastPromotion == 0 ~ "New",
      YearsSinceLastPromotion >= 3 ~ "Passed Over",
      TRUE ~ "On Track")
)
```

d. Years at Company is a skewed distribution variable, peaking early, around 6 years, and trailing off to the max of 40.

Years Since Last Promotion is another skewed variable, with a peak at 1 year and trailing off the the max of 15.

Promotion Status needs better clarity as most employees are in the New group with less in On Track and less again in Passed Over

Education displays a “normal” distribution, with the majority of employees having a Bachelor’s Degree, and reducing from that mode.

Business travel has a large skew:

71% of employees that Rarely Travel

19% of employees Travel Frequently

10% of employees do NOT Travel

Work-Life balance appears healthy:

61% Rate themselves at 3 in work-Life Balance

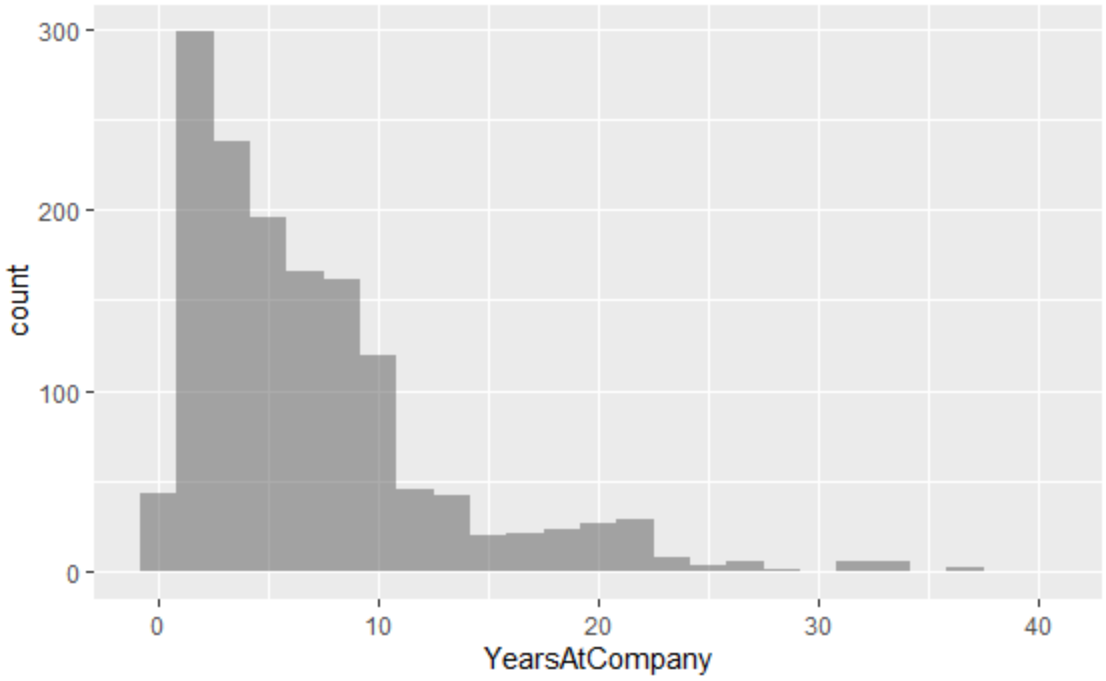
23% Rate themselves at 2 in work-Life Balance

10% Rate themselves at 4 in work-Life Balance

5% Rate themselves at 1 in work-Life Balance

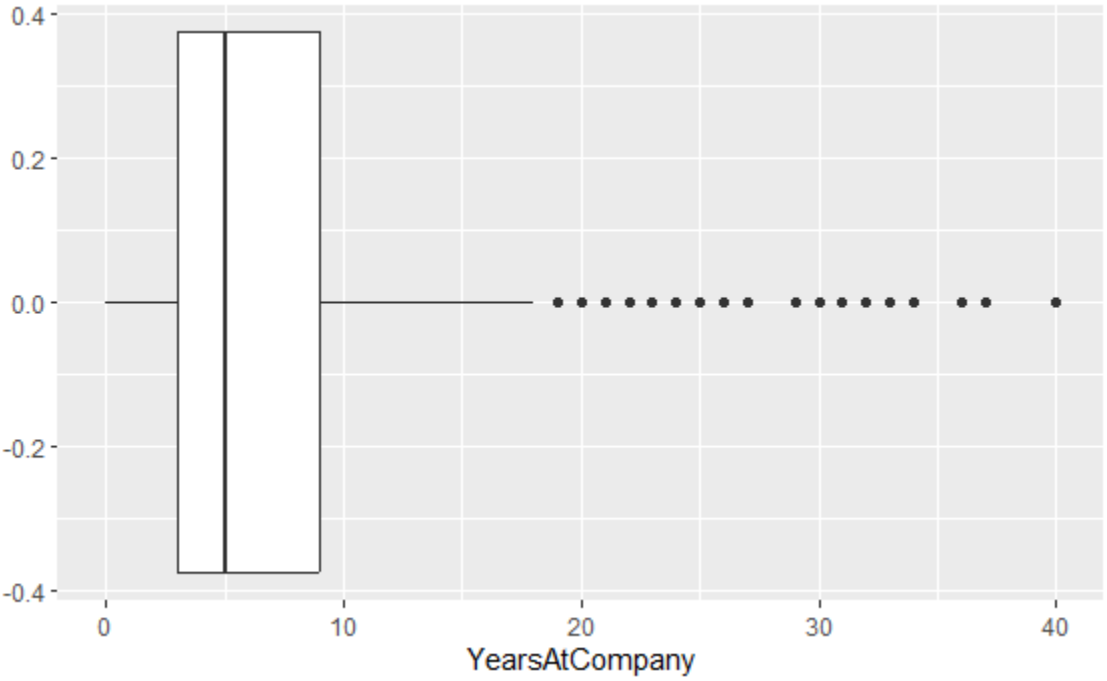
Hide

```
# Did not get any reportable data from str() function
gf_histogram(~YearsAtCompany, data = IBMdata)
```



Hide

```
gf_boxplot(.~YearsAtCompany, data = IBMdata)
```



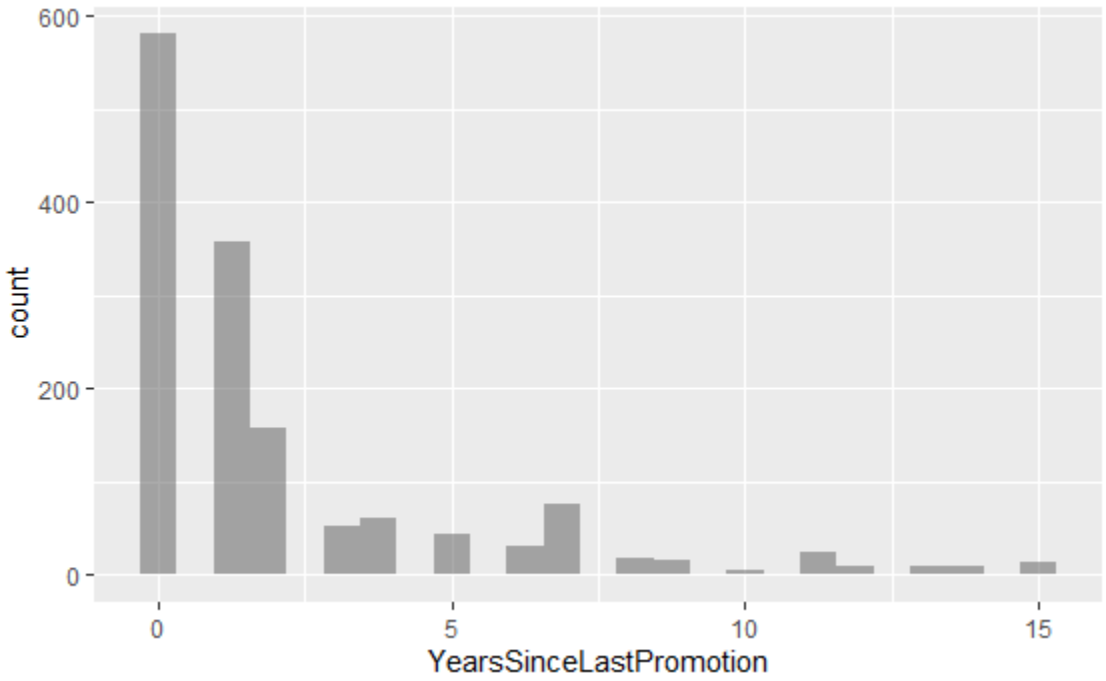
Hide

```
summary(IBMdata$YearsAtCompany)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   3.000   5.000   7.008   9.000  40.000 
```

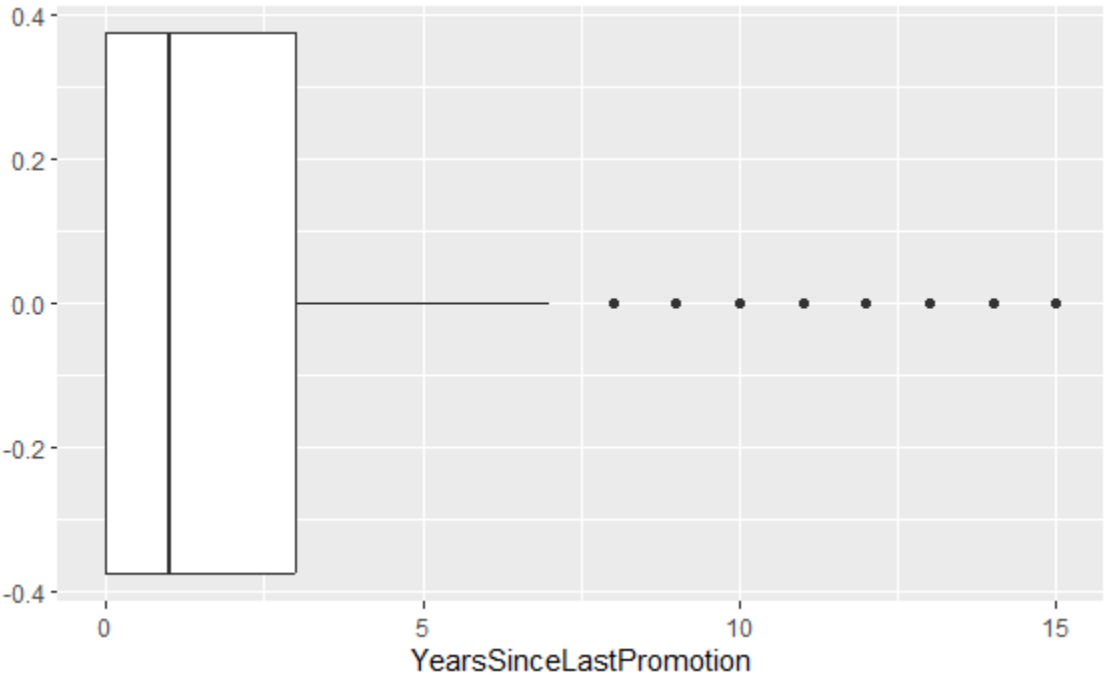
Hide

```
gf_histogram(~YearsSinceLastPromotion, data = IBMdata)
```



Hide

```
gf_boxplot(.~YearsSinceLastPromotion, data = IBMdata)
```



Hide

```
summary(IBMdata$YearsSinceLastPromotion)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   2.188   3.000  15.000
```

Hide

```
gf_bar(~PromotionStatus, data = IBMdata)
```

Hide

```
promotion_status_counts <- table(IBMdata$PromotionStatus)
promotion_status_proportions <- prop.table(promotion_status_counts)
promotion_summary <- data.frame(
  PromotionStatus = names(promotion_status_counts),
  Count = as.numeric(promotion_status_counts),
  Proportion = as.numeric(promotion_status_proportions)
)
print(promotion_summary)
```

PromotionStatus <chr>	Count <dbl>	Proportion <dbl>
New	581	0.3952381
On Track	516	0.3510204
Passed Over	373	0.2537415

3 rows

Hide

```
#(stackoverflow.com, n.d.)

summary(IBMdata$Education)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   2.000   3.000   2.913   4.000   5.000 
```

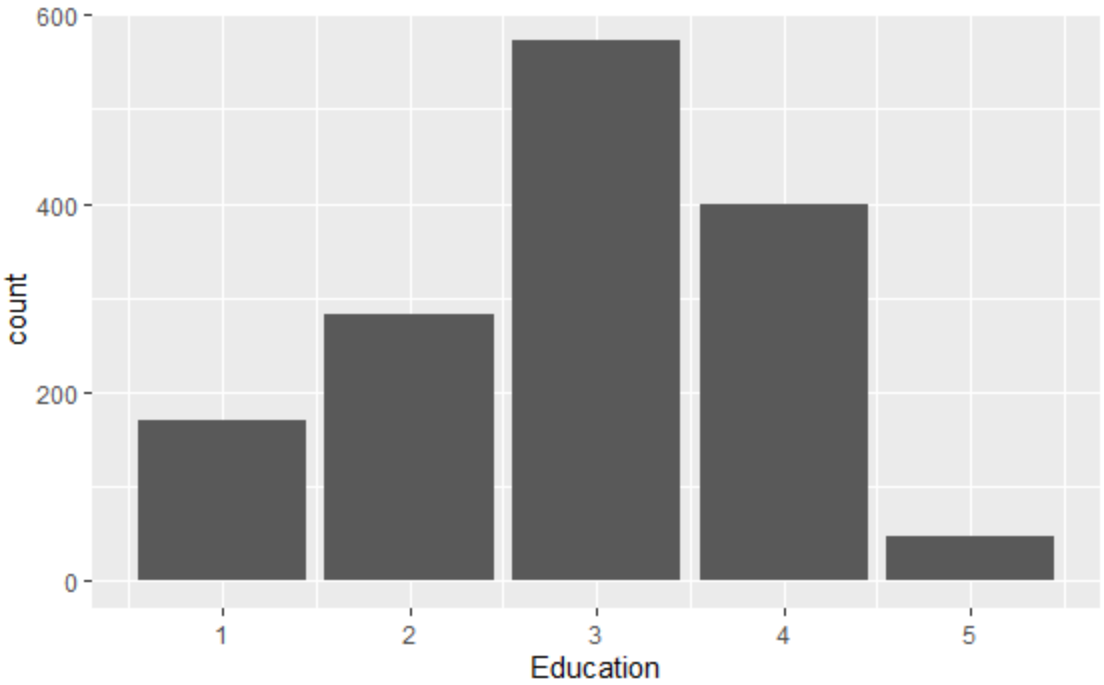
Hide

```
tally(IBMdata$Education)
```

```
X
  1   2   3   4   5 
170 282 572 398  48 
```

Hide

```
gf_bar(~Education, data = IBMdata)
```



Hide

```
summary(IBMdata$BusinessTravel)
```

```
   Length     Class      Mode 
     1470 character character 
```

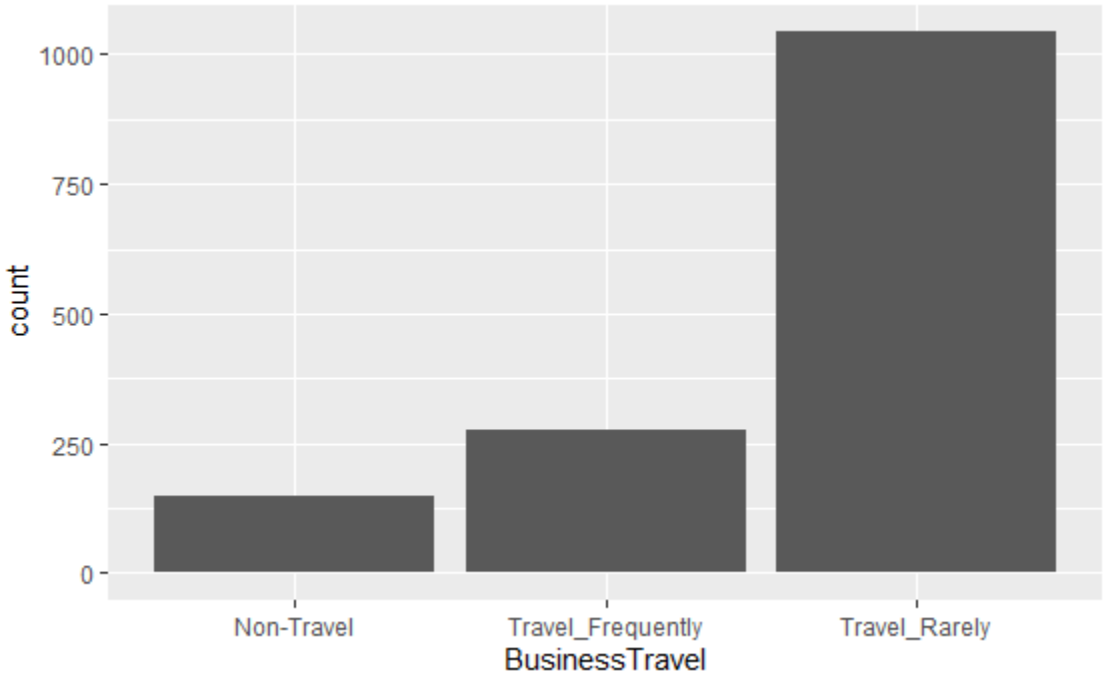
Hide

```
tally(IBMdata$BusinessTravel)
```

```
X
       Non-Travel Travel_Frequently     Travel_Rarely
              150               277              1043
```

Hide

```
gf_bar(~BusinessTravel, data = IBMdata)
```



Hide

```
summary(IBMdata$WorkLifeBalance)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   2.761   3.000   4.000
```

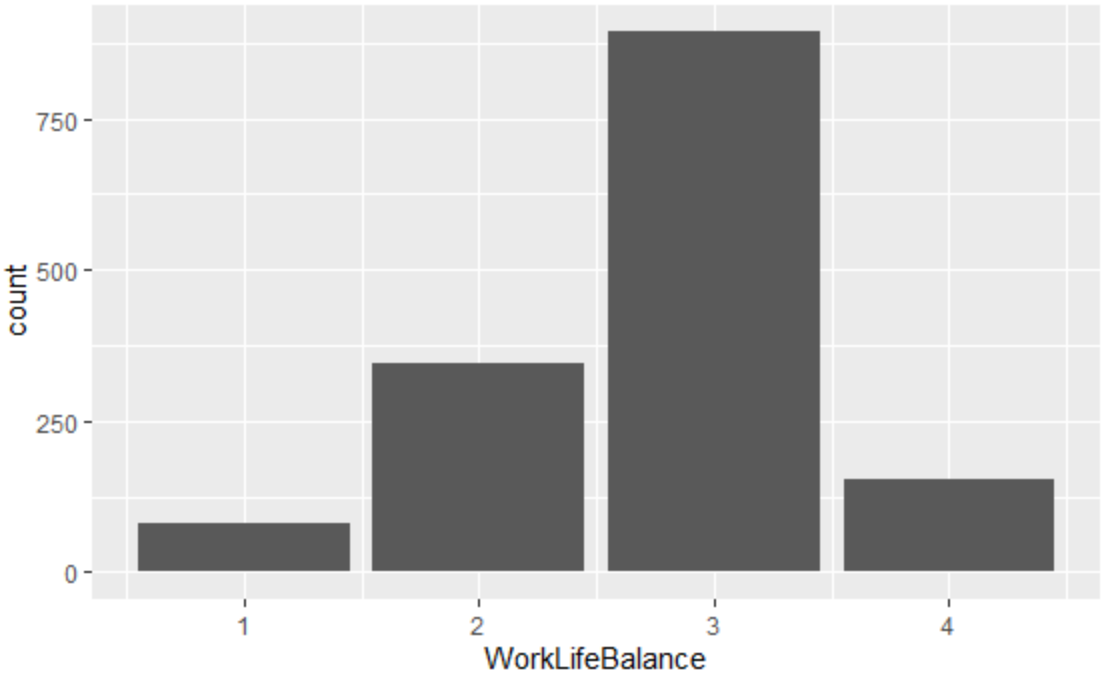
Hide

```
tally(IBMdata$WorkLifeBalance)
```

```
X
  1   2   3   4
 80 344 893 153
```

Hide

```
gf_bar(~WorkLifeBalance, data = IBMdata)
```



e. Looking at the stats for Years at Company, the outliers are above 18, calculated by Q3+1.5*IQR. The code removed 104 rows.

Calculated max for Years Since Promotion the same way, (3+1.5*2), removed 183 rows from IBMdata.

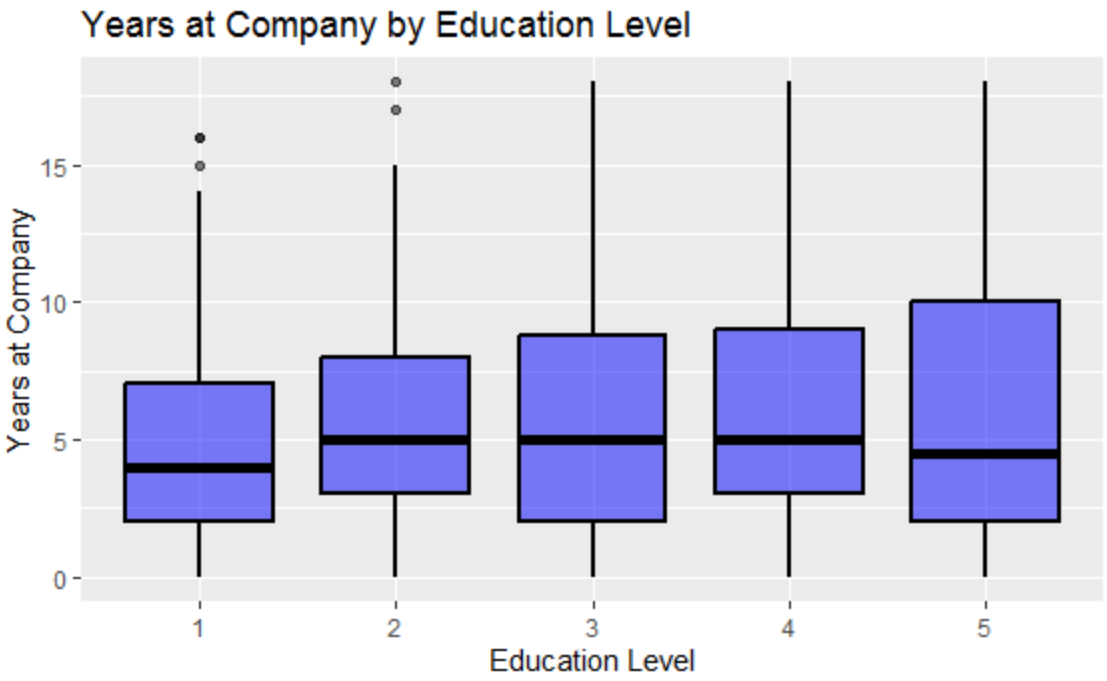
Hide

```
Clean_Years <- filter(IBMdata, YearsAtCompany <= 18)
Promotion_fil <- filter(IBMdata, YearsSinceLastPromotion <= 6)
```

Part3: Visualizations

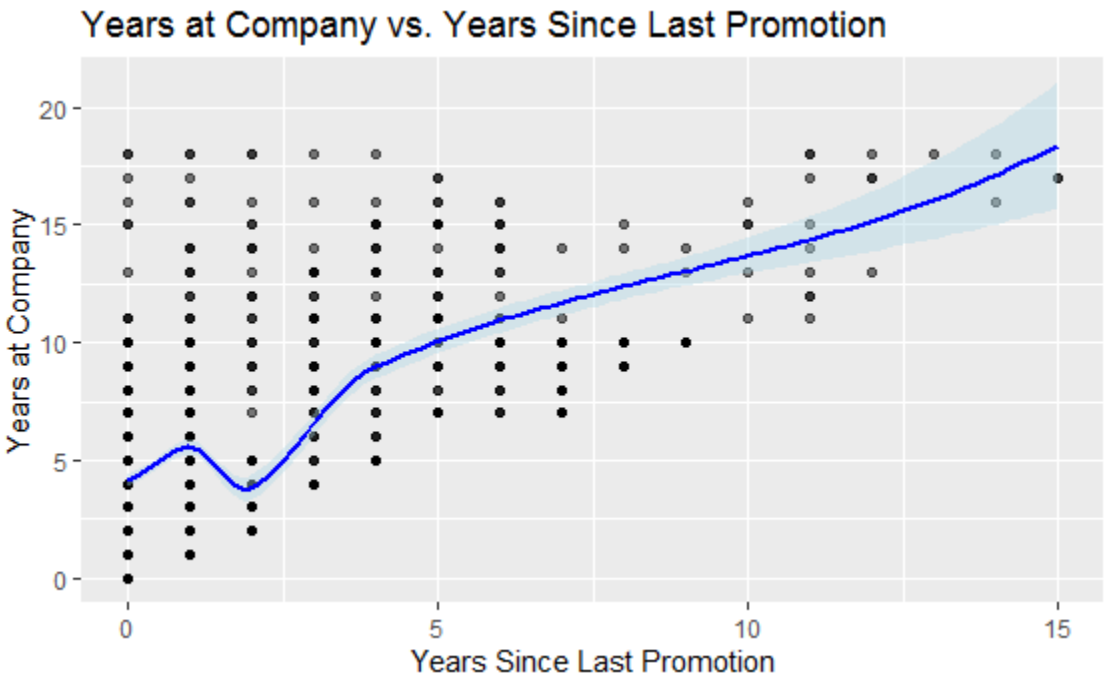
Hide

```
CleanYears$Education <- as.factor(CleanYears$Education)
ggplot(CleanYears,
  aes(x = Education, y = YearsAtCompany)) +
  geom_boxplot(fill = "blue", color = "black",
    size = 1, alpha = 0.5) +
    labs(title = "Years at Company by Education Level",
    x = "Education Level",
    y = "Years at Company")
```



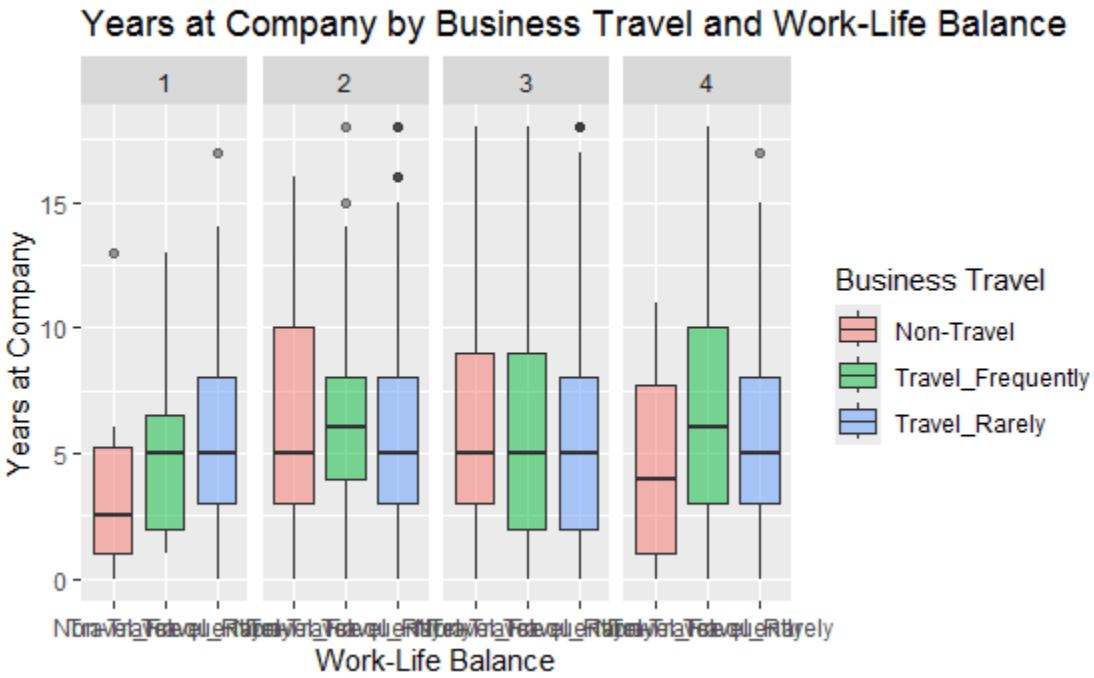
Hide

```
ggplot(CleanYears, aes(x = YearsSinceLastPromotion,
  y = YearsAtCompany)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", color = "blue", fill = "lightblue",
  se = TRUE) +
  labs(title = "Years at Company vs. Years Since Last Promotion",
  x = "Years Since Last Promotion", y = "Years at Company")
```



Hide

```
ggplot(CleanYears, aes(x = BusinessTravel,
  y = YearsAtCompany, fill = BusinessTravel)) +
  geom_boxplot(alpha = 0.5) +
  facet_grid(~ WorkLifeBalance) +
  labs(
  title = "Years at Company by Business Travel and Work-Life Balance",
  x = "Work-Life Balance", y = "Years at Company",
  fill = "Business Travel")    # (stackoverflow.com, n.d.)
```



Part4: Results a. In Chart1 “Years at Company by Education Level,” there is a small amount of deviation in how long an employee stays at the company based on their level of Education (Solid black lines). There is also evidence that employees stay longer, the more education they have (height of the boxes grow from 1-5). Finally, there are a few employees in group 1 and two that have little education, but have been at the company a very long time.

In Chart2 “Years at Company vs. Years Since Last Promotion,” shows a possible hierarchical business model. There appear to be a select group of people that are continuously promoted every 2-3 years, but a majority of the employees are hired to do specific work, with little opportunity for promotion, the best fit line is fairly strait after 4 years, meaning the longer an employee is there the longer since a promotion (i.e. no promotion)

In Chart3 “Years at Company by Business Travel and Work-Life Balance,” there is some differentiation, but it is hard to say that it could be attributed to travel. Except for the Rarely Travel group, the within group differences and across group differences are inconsistent.

b. Chart 1: Roughly even distribution, consistent accross all variables, except for the low outliers. Center and spread are

Chart 2: Good shape, distributed well. spread makes sense, the loess line indicates the unusual nature of promotions previously noted.

Chart 3: The Travel Rarely group is consistent across all work-Life groups. Similar min, Q1, Mean, Q3, max,

spread and outliers,

c. Education has a small affect on longevity. Years Since Promotion does NOT appear to affect the length an employee stays. Business travel combined with Work-Life balance do NOT interact to influence employee retention.

d. It appears that the population sampled is diverse across a majority of the variables. It seems that data was generated from employee records, Annual Reviews, Surveys, and various managerial systems.

####Generalize inferences from sample to the intended population#### Based on the analysis of the dataset, several key inferences can be drawn that may be generalized to the broader employee population within the company.

Introduction This study aims to understand the factors influencing employee retention within a company. By analyzing data on years at the company, education levels, promotion history, business travel frequency, and work-life balance, we seek to identify patterns and relationships that can inform HR policies and practices.

Method Data were collected from employee records, including variables such as YearsAtCompany, Education, YearsSinceLastPromotion, BusinessTravel, and WorkLifeBalance. The sample included employees with varying tenures and backgrounds, providing a comprehensive overview of the workforce.

Results: **Years at Company by Education Level** A boxplot analysis was conducted to examine the distribution of years at the company across different education levels. The median years at the company varied by education level, with employees holding advanced degrees tending to have longer tenures (see Chart 1).

Years at Company vs. Years Since Last Promotion A scatter plot with a smooth density overlay was used to explore the relationship between years at the company and years since the last promotion. The results indicated a positive trend, suggesting that employees who had not been promoted recently had longer tenures (see Chart 2).

Years at Company by Business Travel and Work-Life Balance Faceted boxplots were used to analyze the interaction between business travel frequency and work-life balance on years at the company. Employees with balanced work-life perceptions and moderate travel requirements tended to have longer tenures (see Chart 3).

Discussion The findings from this analysis reveal several important insights about employee retention:

1. **Education Level and Retention**: Employees with higher education levels, tend to have longer tenures within the company. This suggests that the company values advanced education, which correlates with career stability.

2. **Promotion History and Retention**: There is no discernible relationship between promotion history and tenure. Employees who have not been promoted recently tend to have longer tenures. This highlights a lack of promotion opportunities.

3. **Business Travel, Work-Life Balance, and Retention**: Business travel frequency does not seem to affect overall work-life balance no apparent work-life balance issue issignificantly impacting employee retention.

Conclusion This study has the potential to provide valuable insights into the factors influencing employee retention within the company. By looking at other gathered variable, we should be able to identify factors that do impact retention. Then, HR policies can be tailored to enhance employee satisfaction and reduce turnover.

Part 5: Implications The potential sources of errors in collecting Education data are explored below, other variables that could explain variation, and possible confounding variables for the hypothesis that Education

would impact how lang an employee stays at the company.

1. **Measurement Error** If an employee's highest education level is mistakenly entered.
2. **Sampling Error** If the sample is not representative of the entire employee population, such as excluding recent hires or long-term employees, the results might not generalize.
3. **Mistakes** Data entry mistakes, such as transposing numbers in a field.

Other Variables in the Dataset that should be analyzed:

1. **Job Role** Different job roles might have different typical tenures.
2. **Department** Employees in different departments might experience different retention rates due to varying work environments, management styles, or job stability.

Other Variables that should be added to the Dataset:

1. **Job Satisfaction / Engagement** Employee's satisfaction with their job role and responsibilities. How engaged they are with their work.
2. **Culture / Work Environment** A measure of the overall work environment quality, including factors like safety, coworker relationships, and office facilities.
3. **Management Support / Team Dynamics** Perception of support and encouragement from immediate supervisors. Quality of other interactions and relationships within the team.

Confounding Variables:

1. **Age** Age could be a confounding variable if older employees both have higher education levels and longer tenures simply due to having been in the workforce longer. Not accounting for age could lead to overestimating the effect of education on tenure.
2. **Performance** High-performing employees might both have longer tenures and higher education levels.

Part 6: Discussion The analysis report provides actionable insights into the factors influencing employee retention. By understanding how education, promotion history, business travel, and work-life balance impact years at the company, employers can tailor their HR policies to enhance employee satisfaction and reduce turnover.

The practical significance lies in identifying key areas for intervention to improve retention rates. This is crucial as retaining experienced employees reduces recruitment costs, maintains institutional knowledge, and promotes a positive workplace culture.

Understanding the nature and circumstances influencing these relationships helps employers develop targeted strategies that address specific needs and challenges faced by employees. This holistic approach ensures initiatives are not just well-intentioned but effective in retaining talent and fostering a supportive work environment.