Load Libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(RColorBrewer)
library(cowplot)
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(ggthemes)
```

```
##
## Attaching package: 'ggthemes'

## The following object is masked from 'package:cowplot':
##
##     theme_map
```

Load in the data

```
data = read.csv(file ="https://raw.githubusercontent.com/BivinSadler/MSDS_6306_Doing-Data-Science/Maste
                        stringsAsFactors = T,
                        header = T)
```

Look at the data

```
str(data)
```

```
## 'data.frame':    870 obs. of  36 variables:
##  $ ID                      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age                     : int  32 40 35 32 24 27 41 37 34 34 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 2 3 2 2 3 3
##  $ DailyRate               : int  117 1308 200 801 567 294 1283 309 1333 653 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 3 2 2 2 3 3 2 ...
##  $ DistanceFromHome        : int  13 14 18 1 2 10 5 10 10 10 ...
##  $ Education               : int  4 3 2 4 1 2 5 4 4 4 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 4 2 3 6 2 4 2 2 6 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  859 1128 1412 2016 1646 733 1448 1105 1055 1597 ...
##  $ EnvironmentSatisfaction : int  2 3 3 3 1 4 2 4 3 4 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 2 2 2 1 1 2 2 1 1 2 ...
##  $ HourlyRate              : int  73 44 60 48 32 32 90 88 87 92 ...
##  $ JobInvolvement          : int  3 2 3 3 3 3 4 2 3 2 ...
##  $ JobLevel                : int  2 5 3 3 1 3 1 2 1 2 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 6 5 8 7 5 7 8 9 1
##  $ JobSatisfaction         : int  4 3 4 4 4 1 3 4 3 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 1 3 3 2 3 1 2 1 2 2 ...
##  $ MonthlyIncome           : int  4403 19626 9362 10422 3760 8793 2127 6694 2220 5063 ...
##  $ MonthlyRate             : int  9250 17544 19944 24032 17218 4809 5561 24223 18410 15332 ...
##  $ NumCompaniesWorked      : int  2 1 2 1 1 1 2 2 1 1 ...
##  $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 2 2 1 ...
##  $ PercentSalaryHike       : int  11 14 11 19 13 21 12 14 19 14 ...
##  $ PerformanceRating       : int  3 3 3 3 3 4 3 3 3 3 ...
##  $ RelationshipSatisfaction: int  3 1 3 3 3 3 1 3 4 2 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  1 0 0 2 0 2 0 3 1 1 ...
##  $ TotalWorkingYears       : int  8 21 10 14 6 9 7 8 1 8 ...
##  $ TrainingTimesLastYear   : int  3 2 2 3 2 4 5 5 2 3 ...
##  $ WorkLifeBalance         : int  2 4 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : int  5 20 2 14 6 9 4 1 1 8 ...
##  $ YearsInCurrentRole      : int  2 7 2 10 3 7 2 0 1 2 ...
##  $ YearsSinceLastPromotion : int  0 4 2 5 1 1 0 0 0 7 ...
##  $ YearsWithCurrManager    : int  3 9 2 7 3 7 3 0 0 7 ...
```

```
head(data)
```

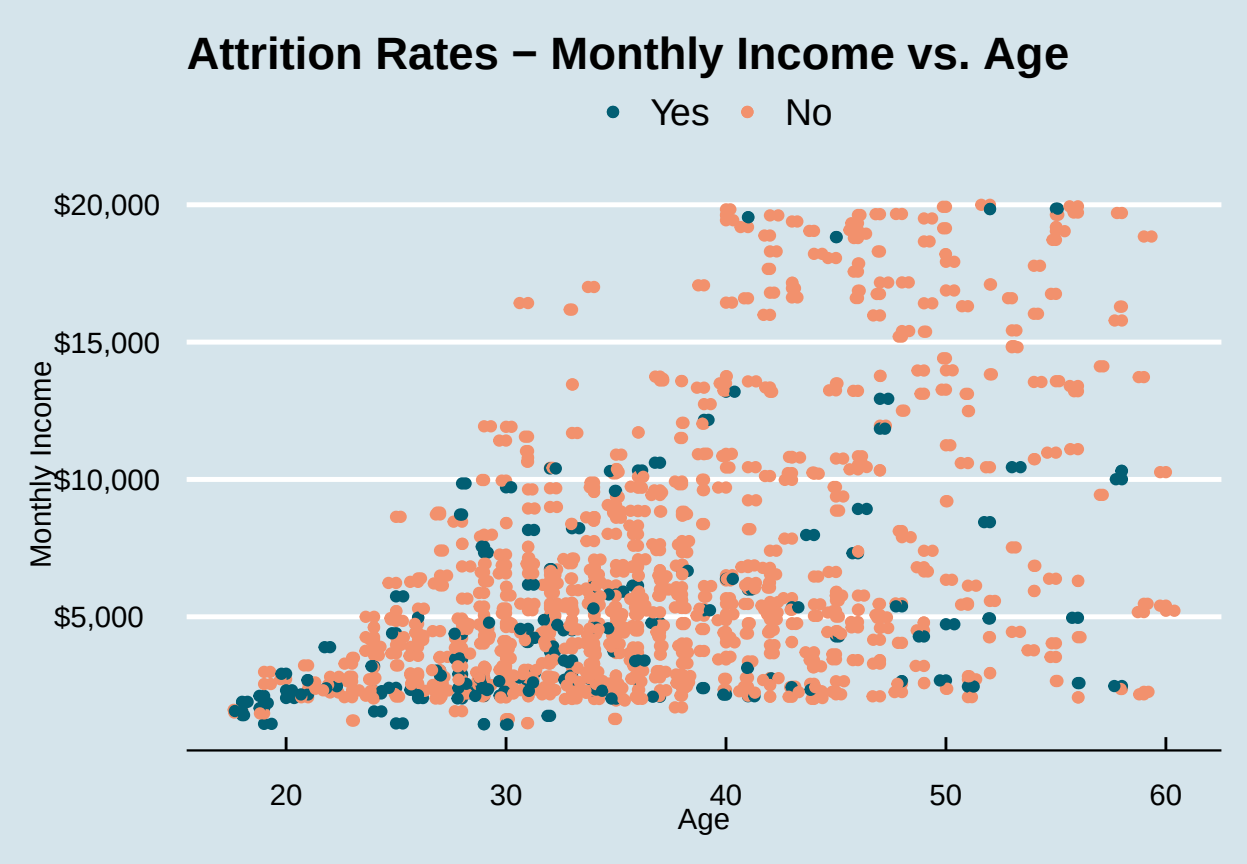
```
##   ID Age Attrition    BusinessTravel DailyRate              Department
## 1  1  32        No     Travel_Rarely       117                   Sales
## 2  2  40        No     Travel_Rarely      1308 Research & Development
## 3  3  35        No Travel_Frequently       200 Research & Development
## 4  4  32        No     Travel_Rarely       801                   Sales
## 5  5  24        No Travel_Frequently       567 Research & Development
## 6  6  27        No Travel_Frequently       294 Research & Development
##   DistanceFromHome Education   EducationField EmployeeCount EmployeeNumber
## 1               13         4    Life Sciences             1            859
## 2               14         3          Medical             1           1128
```

```
## 3               18         2    Life Sciences             1           1412
## 4                1         4       Marketing              1           2016
## 5                2         1 Technical Degree             1           1646
## 6               10         2    Life Sciences             1            733
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       2   Male         73              3        2
## 2                       3   Male         44              2        5
## 3                       3   Male         60              3        3
## 4                       3 Female         48              3        3
## 5                       1 Female         32              3        1
## 6                       4   Male         32              3        3
##                 JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1       Sales Executive               4      Divorced          4403
## 2     Research Director               3        Single         19626
## 3 Manufacturing Director               4        Single          9362
## 4       Sales Executive               4       Married         10422
## 5    Research Scientist               4        Single          3760
## 6 Manufacturing Director               1      Divorced          8793
##   MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 1        9250                 2      Y       No                11
## 2       17544                 1      Y       No                14
## 3       19944                 2      Y       No                11
## 4       24032                 1      Y       No                19
## 5       17218                 1      Y      Yes                13
## 6        4809                 1      Y       No                21
##   PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
## 1                 3                        3           80                1
## 2                 3                        1           80                0
## 3                 3                        3           80                0
## 4                 3                        3           80                2
## 5                 3                        3           80                0
## 6                 4                        3           80                2
##   TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
## 1                 8                     3               2             5
## 2                21                     2               4            20
## 3                10                     2               3             2
## 4                14                     3               3            14
## 5                 6                     2               3             6
## 6                 9                     4               2             9
##   YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
## 1                  2                       0                    3
## 2                  7                       4                    9
## 3                  2                       2                    2
## 4                 10                       5                    7
## 5                  3                       1                    3
## 6                  7                       1                    7
```

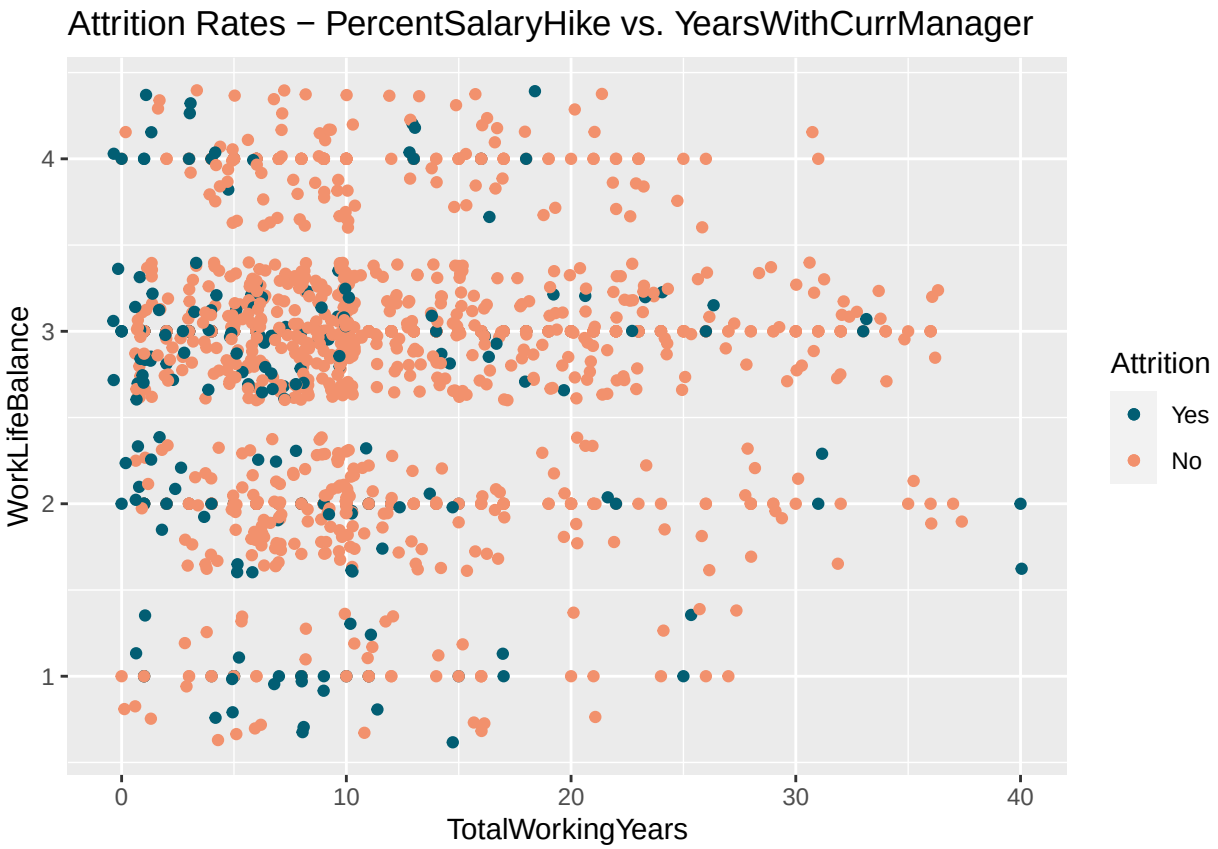
```
##Looking for relationships, SCATTER PLOTS
data %>% ggplot(aes(x=Age, y=MonthlyIncome, color=Attrition)) +
  geom_point() +
  geom_jitter() +
  scale_color_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  ggtitle("Attrition Rates - Monthly Income vs. Age") +
  scale_y_continuous(labels = scales::comma) +
```

```
  scale_y_continuous(labels=scales::dollar_format()) +
  labs(y="Monthly Income") +
  theme_economist() +
  theme(legend.title = element_blank())
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```



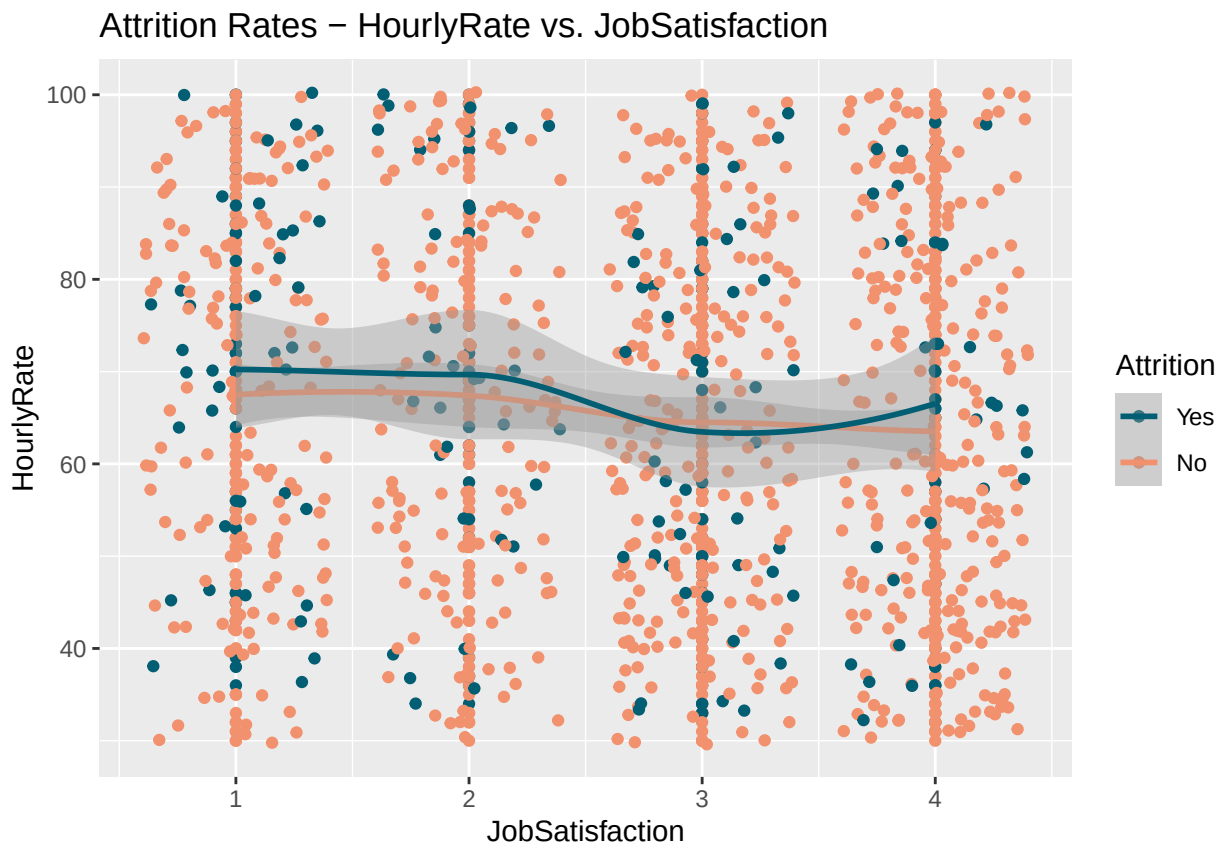

```
#TotalWorkingYears vs WorkLifeBalance, in regards to Attrition
data %>% ggplot(aes(x=TotalWorkingYears, y=WorkLifeBalance, color = Attrition)) +
  geom_point() +
  geom_jitter() +
  scale_color_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  ggtitle("Attrition Rates - PercentSalaryHike vs. YearsWithCurrManager")
```




```
#JobSatisfaction vs HourlyRate, in regards to Attrition
data %>% ggplot(aes(x=JobSatisfaction, y=HourlyRate, color=Attrition)) +
  geom_point() +
  geom_jitter() +
  geom_smooth() +
  scale_color_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  ggtitle("Attrition Rates - HourlyRate vs. JobSatisfaction")
```

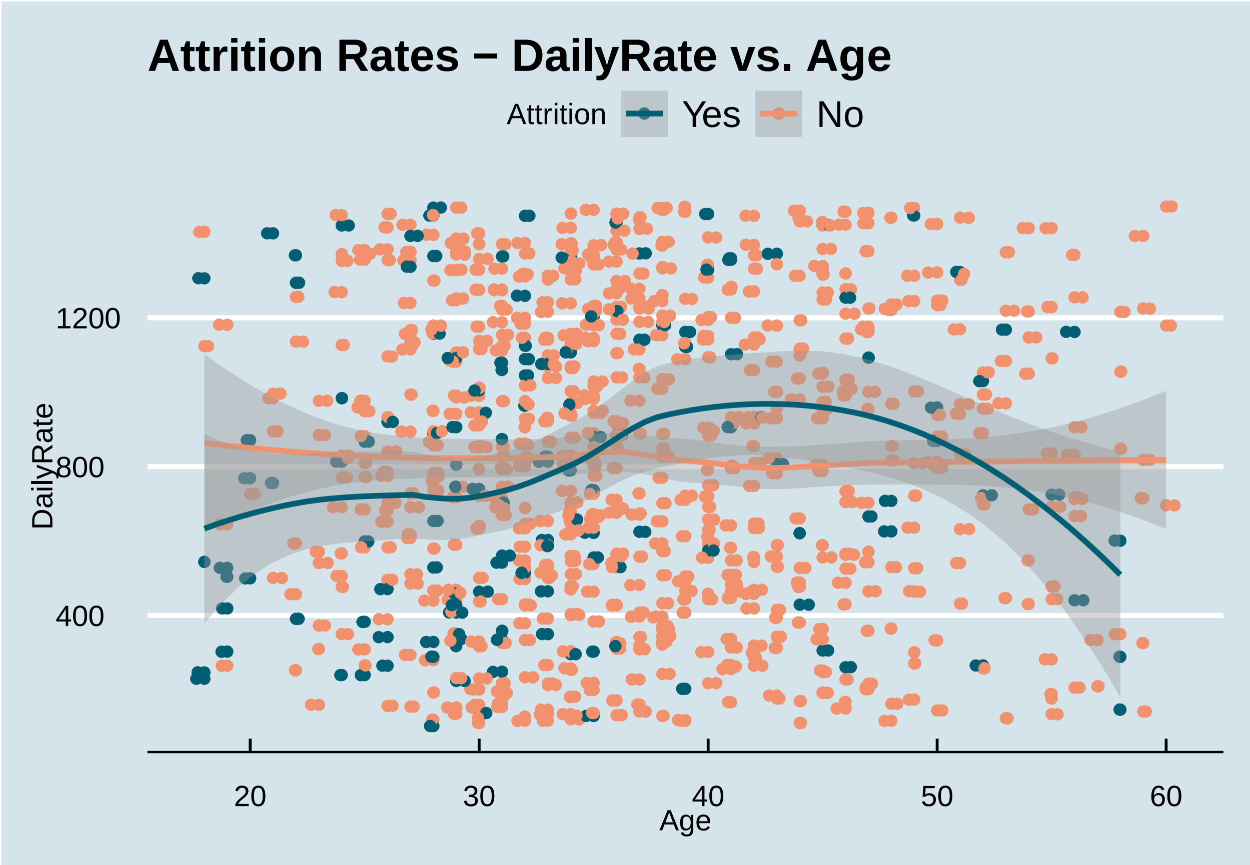
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 4.015

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 2.015

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 2.2876e-015

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## 4.015
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 2.015

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 2.2876e-015

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 0.985

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 2.015

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 5.0559e-017

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## 0.985

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 2.015

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 5.0559e-017

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 1
```

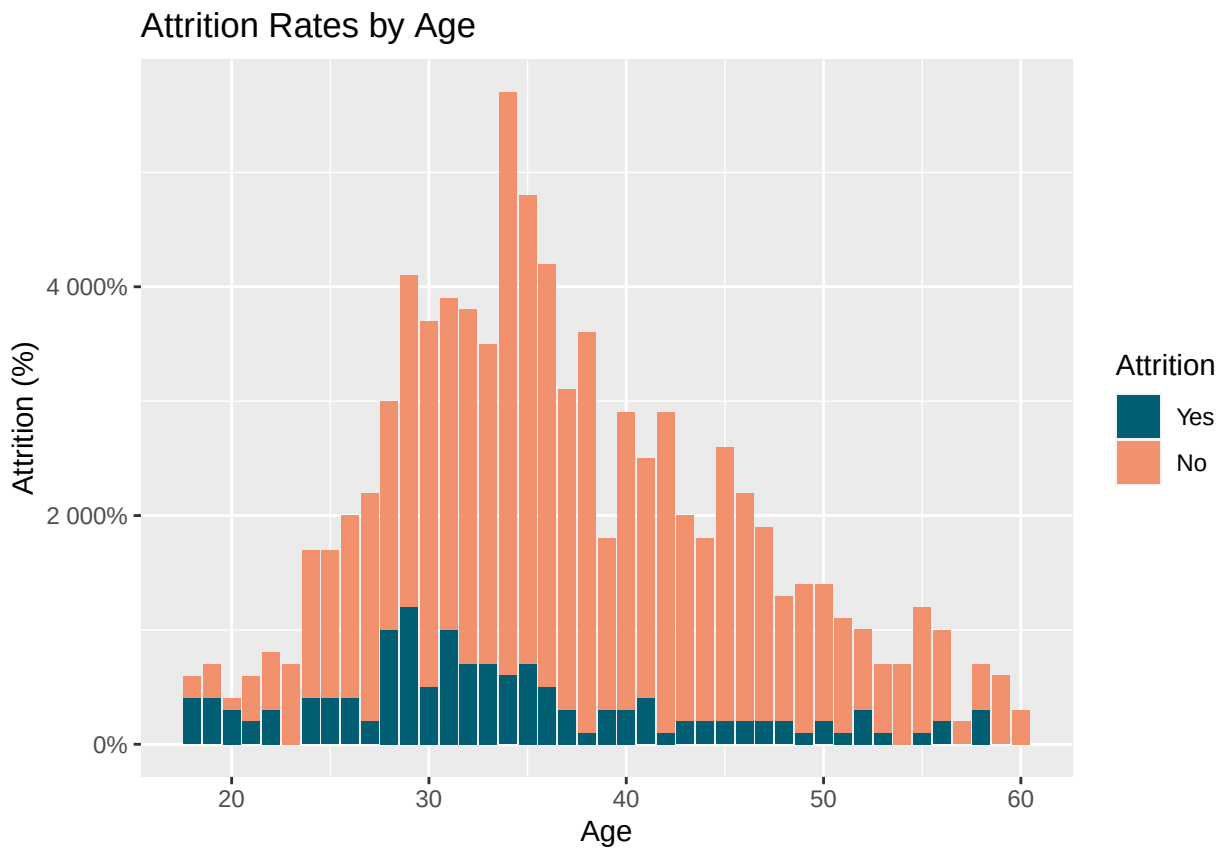



```
#Age vs DailyRate, in regards to Attrition
data %>% ggplot(aes(x=Age, y=DailyRate, color=Attrition)) +
  geom_point() +
  geom_jitter() +
  geom_smooth() +
  scale_color_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  ggtitle("Attrition Rates - DailyRate vs. Age") +
  theme_economist()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

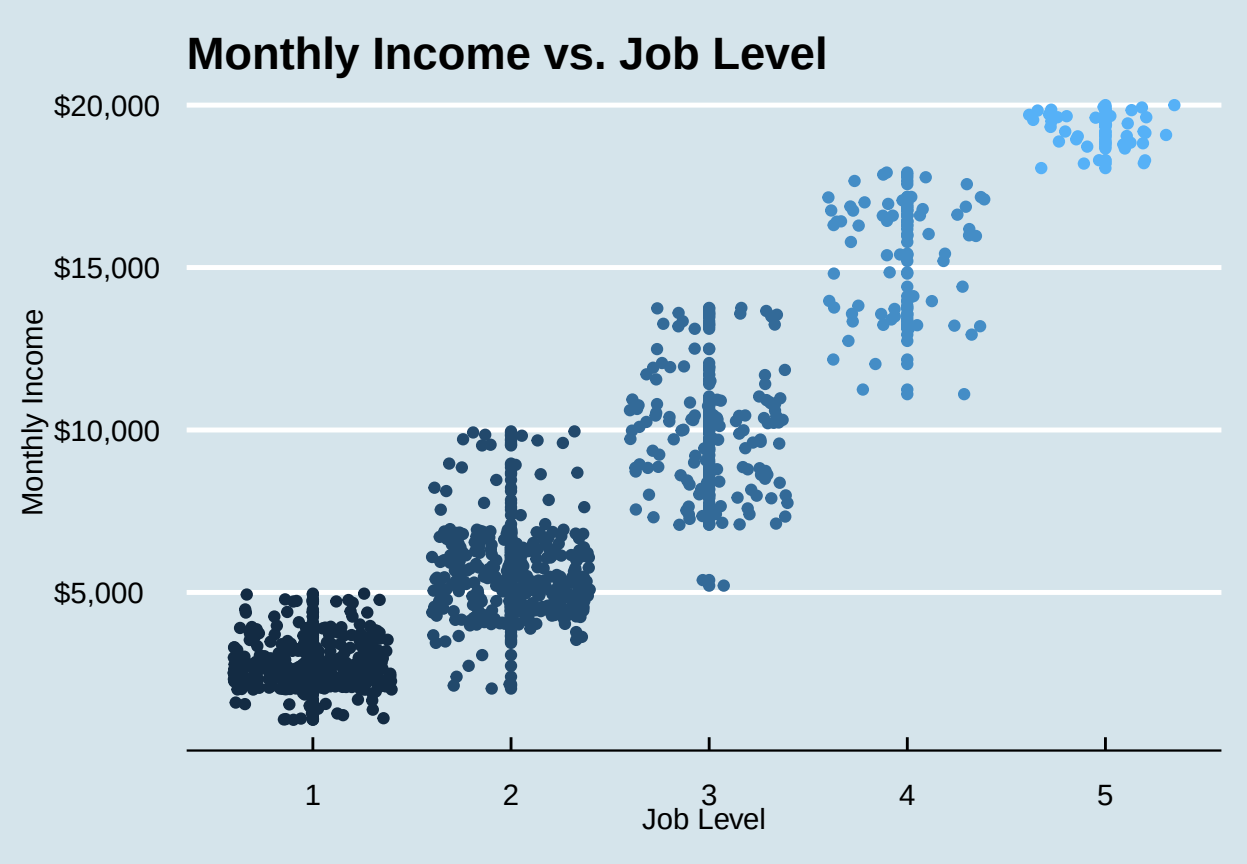



```
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

##Looking for relationships, BAR GRAPHS
data %>% ggplot(aes(x=Age, fill=Attrition)) +
  geom_bar() +
  scale_fill_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  ggtitle("Attrition Rates by Age") +
  scale_y_continuous(labels = scales::percent) +
  labs(y="Attrition (%)")
```

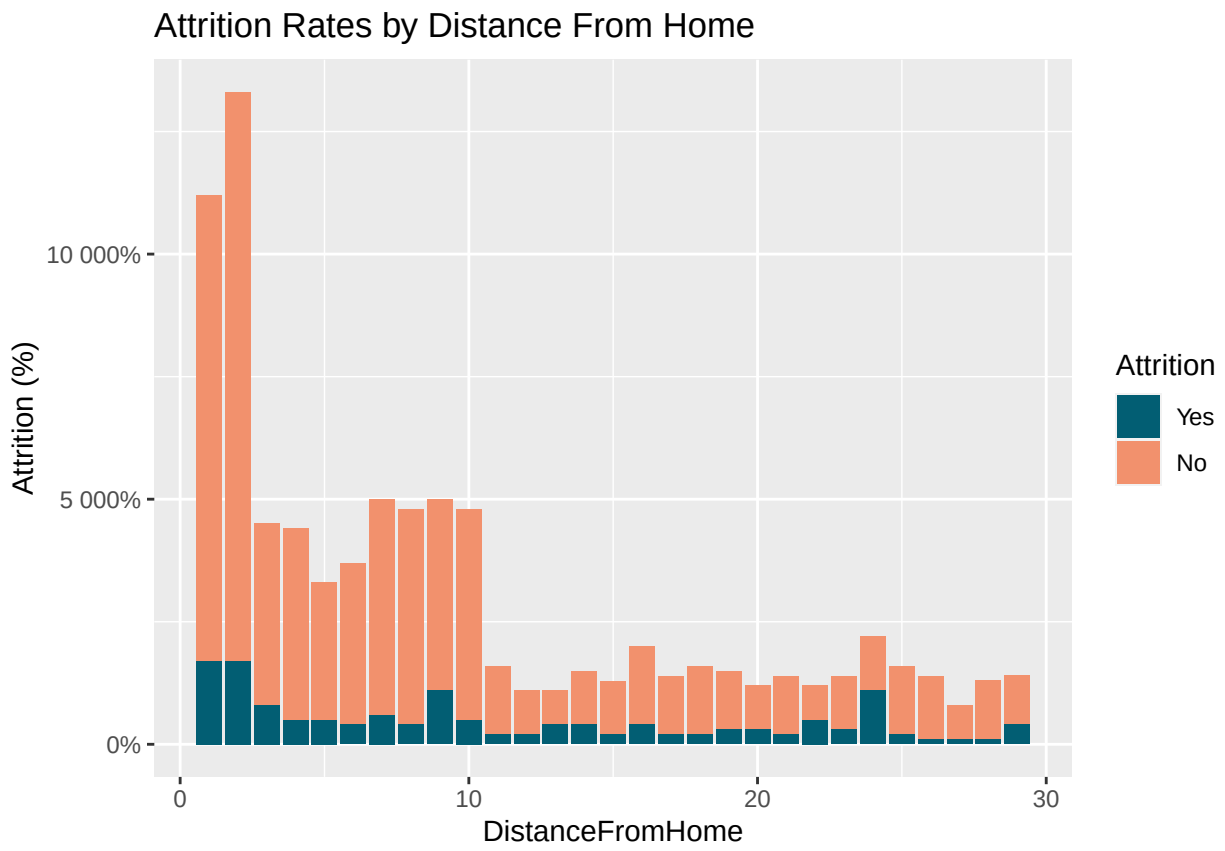



```
#JobLevel vs MonthlyIncome
data %>% ggplot(aes(x=JobLevel, y=MonthlyIncome, color=JobLevel)) +
  geom_point() +
  geom_jitter() +
  ggtitle("Monthly Income vs. Job Level") +
  labs(y="Monthly Income", x="Job Level") +
  scale_y_continuous(labels = scales::comma)+
  scale_y_continuous(labels=scales::dollar_format()) +
  theme_economist() +
  theme(legend.position = "None", axis.title.y=element_text(vjust=1.8))
```

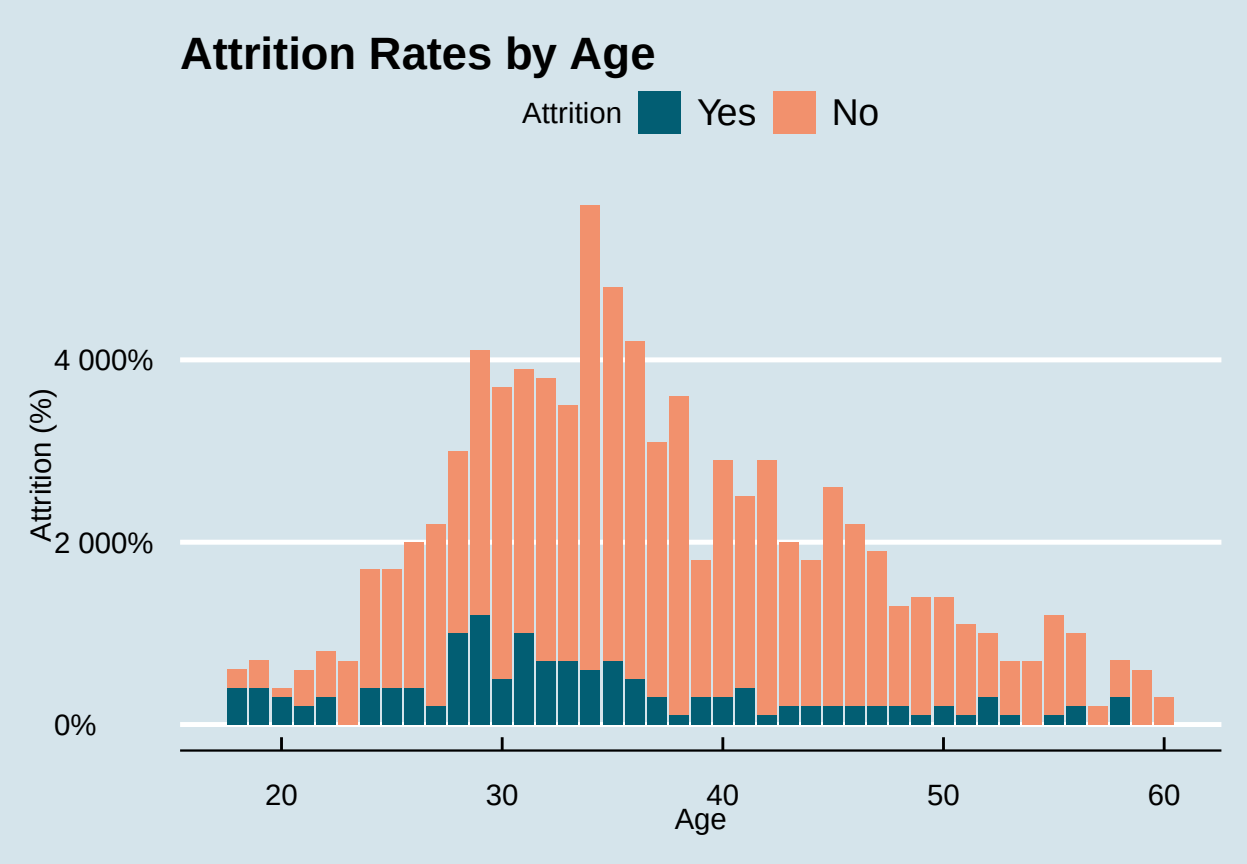
```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```



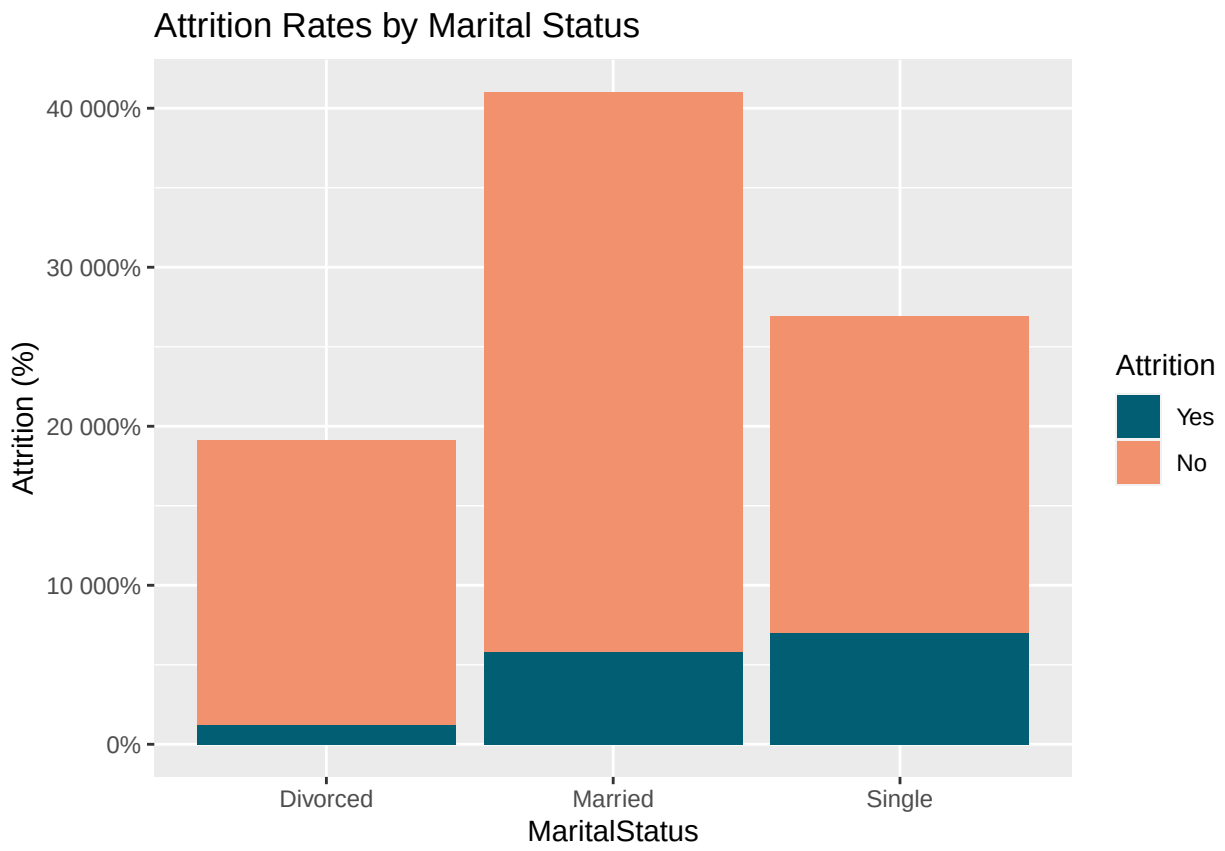

```
#DistanceFromHome in regards to Attrition
data %>% ggplot(aes(x=DistanceFromHome, fill=Attrition)) +
  geom_bar() +
  scale_fill_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  ggtitle("Attrition Rates by Distance From Home") +
  scale_y_continuous(labels = scales::percent) +
  labs(y="Attrition (%)")
```



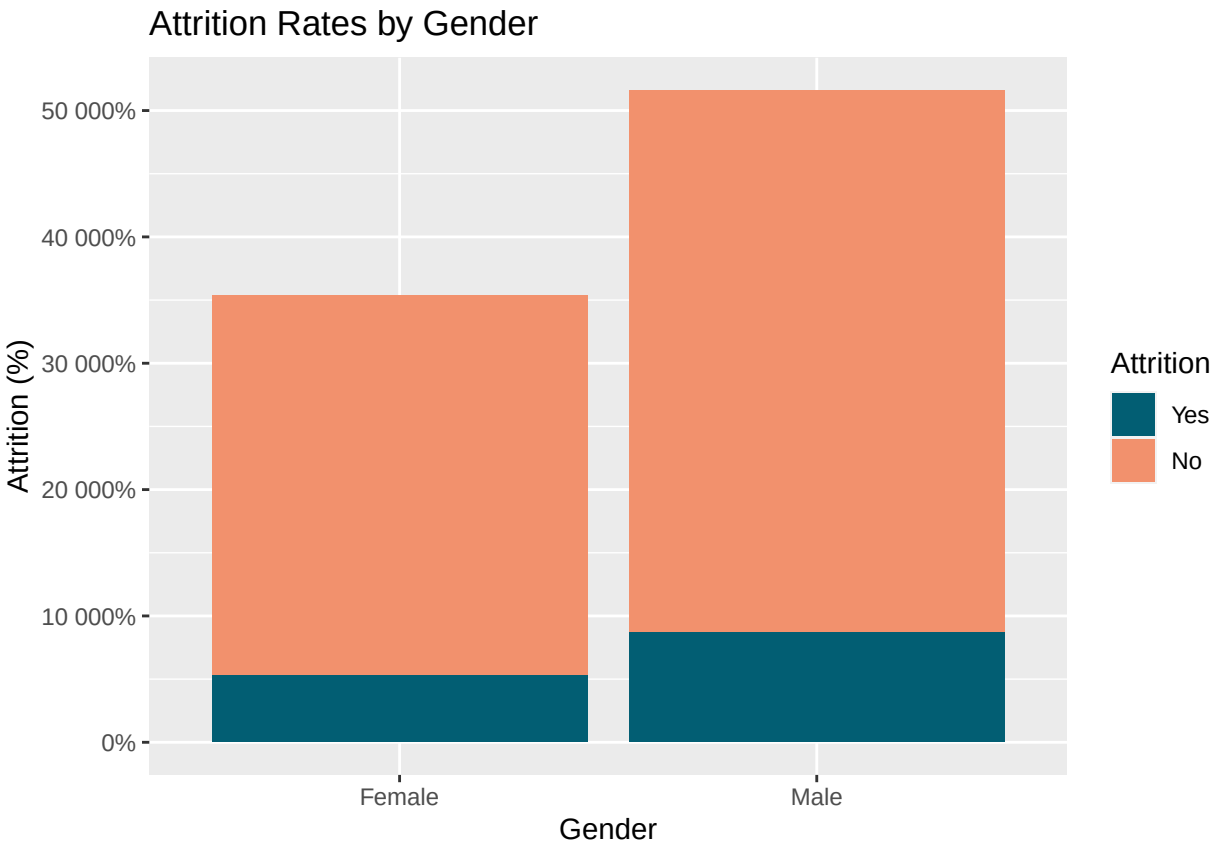

```
data %>% ggplot(aes(x=Age, fill=Attrition)) +
  geom_bar() +
  scale_fill_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  ggtitle("Attrition Rates by Age") +
  scale_y_continuous(labels = scales::percent) +
  labs(y="Attrition (%)")  +
  theme_economist()
```



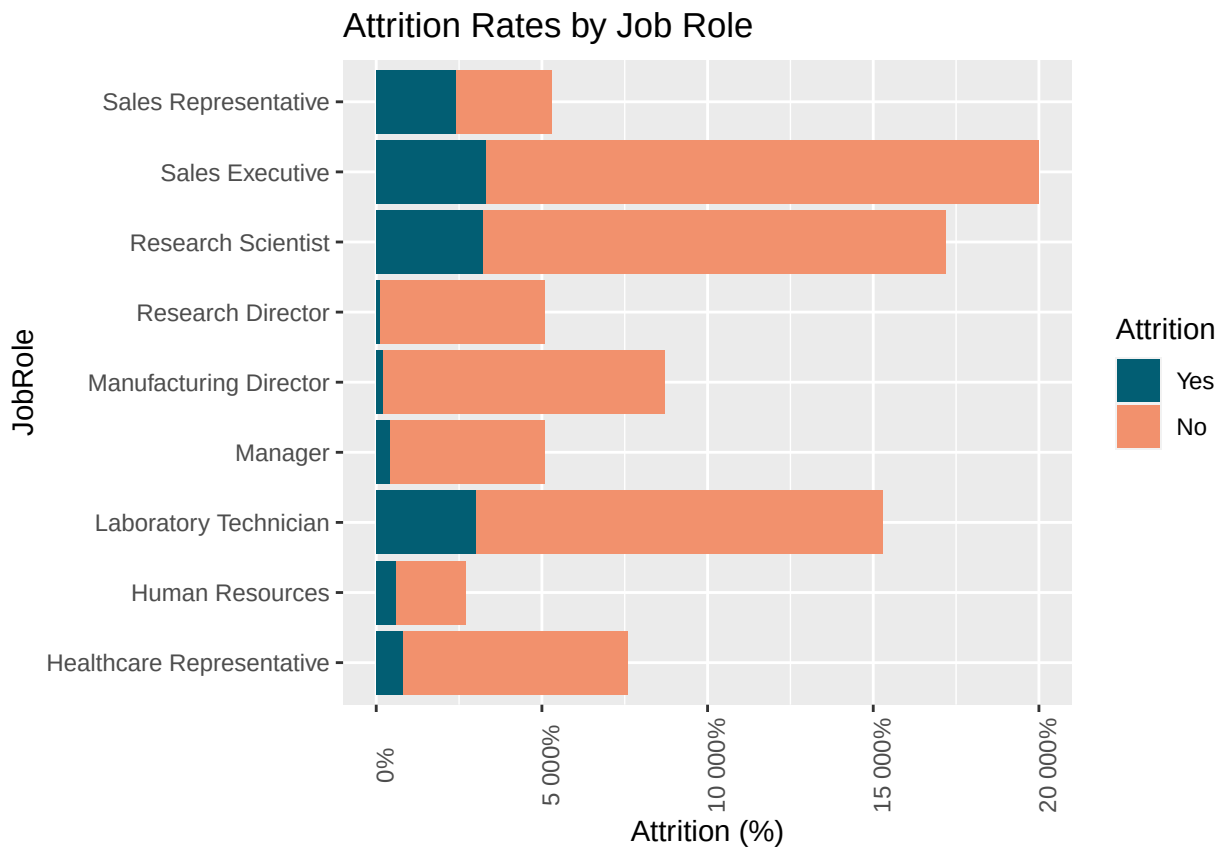

```
#MaritalStatus in regards to Attrition
data %>% ggplot(aes(x=MaritalStatus, fill=Attrition)) +
  geom_bar() +
  scale_fill_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  ggtitle("Attrition Rates by Marital Status") +
  scale_y_continuous(labels = scales::percent) +
  labs(y="Attrition (%)")
```




```
#Gender in regards to Attrition
data %>% ggplot(aes(x=Gender, fill=Attrition)) +
  geom_bar() +
  scale_fill_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  ggtitle("Attrition Rates by Gender") +
  scale_y_continuous(labels = scales::percent) +
  labs(y="Attrition (%)")
```



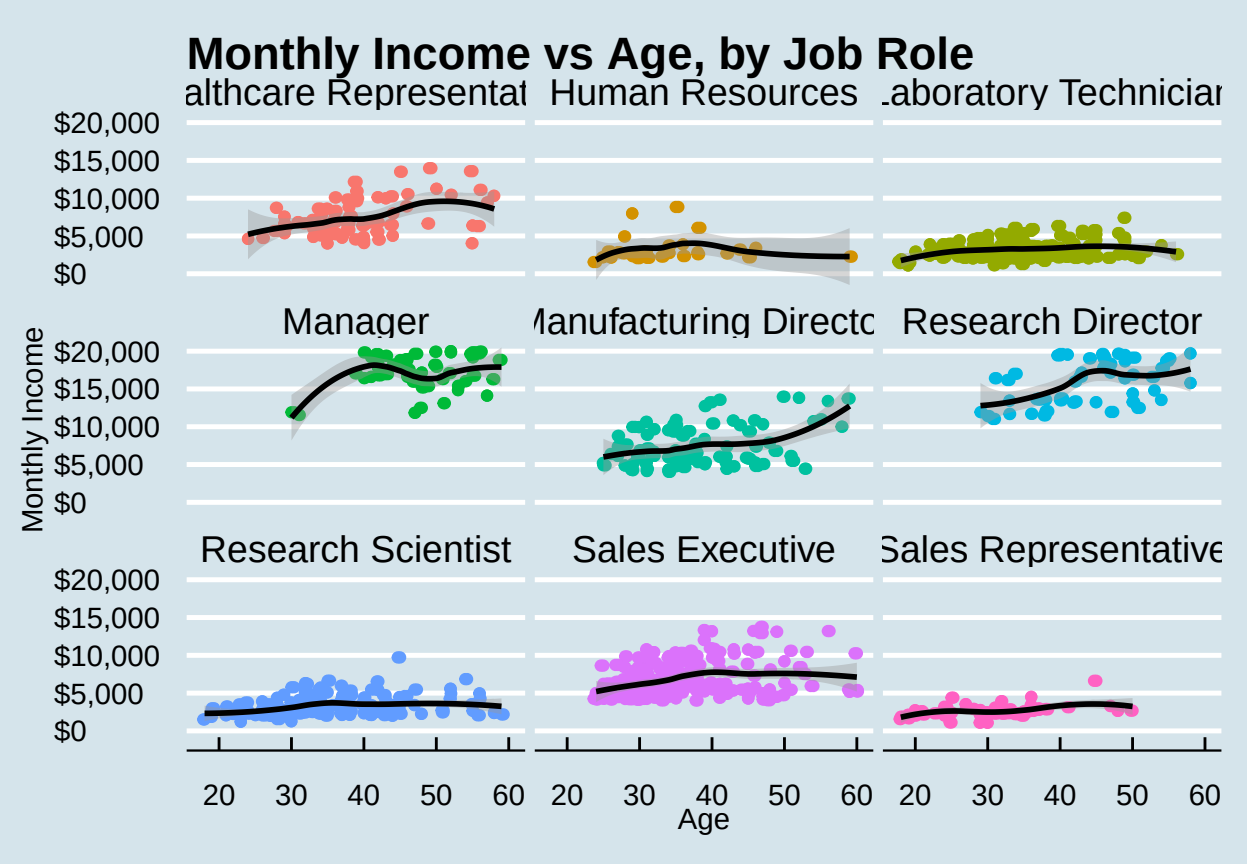

```
#JobRole in regards to Attrition
data %>% ggplot(aes(x=JobRole, fill=Attrition)) +
  geom_bar() +
  theme(axis.text.x=element_text(angle=90)) +
  scale_fill_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  ggtitle("Attrition Rates by Job Role") +
  scale_y_continuous(labels = scales::percent) +
  labs(y="Attrition (%)") +
  coord_flip()
```



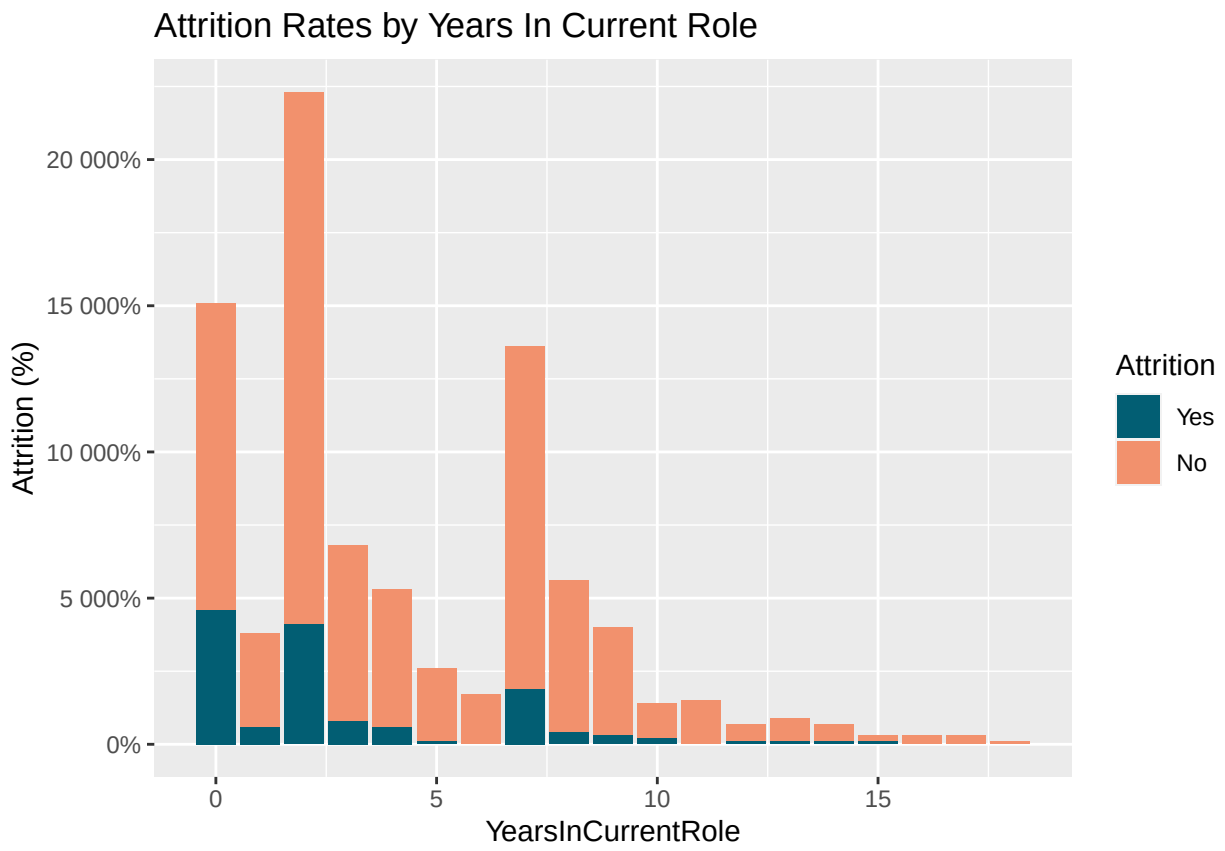

```
#Facet wrap of Age vs. MonthlyIncome, by Job Role
data %>% ggplot(aes(x=Age, y=MonthlyIncome, color=JobRole)) +
  geom_point() +
  geom_jitter() +
  geom_smooth(color = "black") +
  facet_wrap(~JobRole) + #facet wrap
  ggtitle("Monthly Income vs Age, by Job Role") +
  labs(y="Monthly Income") +
  scale_y_continuous(labels = scales::comma)+
  scale_y_continuous(labels=scales::dollar_format()) +
  theme_economist() +
  theme(legend.position = "None", axis.title.y=element_text(vjust=1.8))
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```

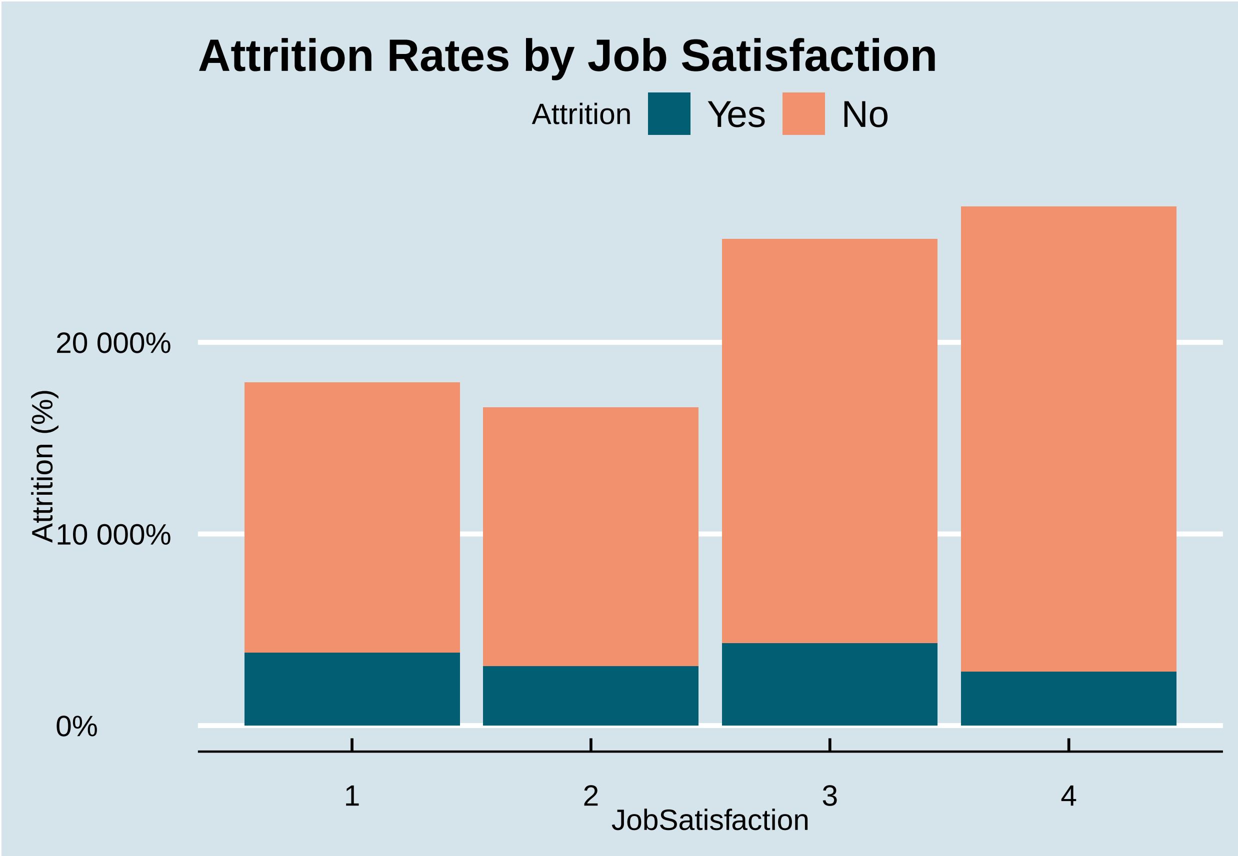
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
#YearsInCurrentRole in relation to Attrition
data %>% ggplot(aes(x=YearsInCurrentRole, fill=Attrition)) +
  geom_bar() +
  scale_fill_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  ggtitle("Attrition Rates by Years In Current Role") +
  scale_y_continuous(labels = scales::percent) +
  labs(y="Attrition (%)")
```

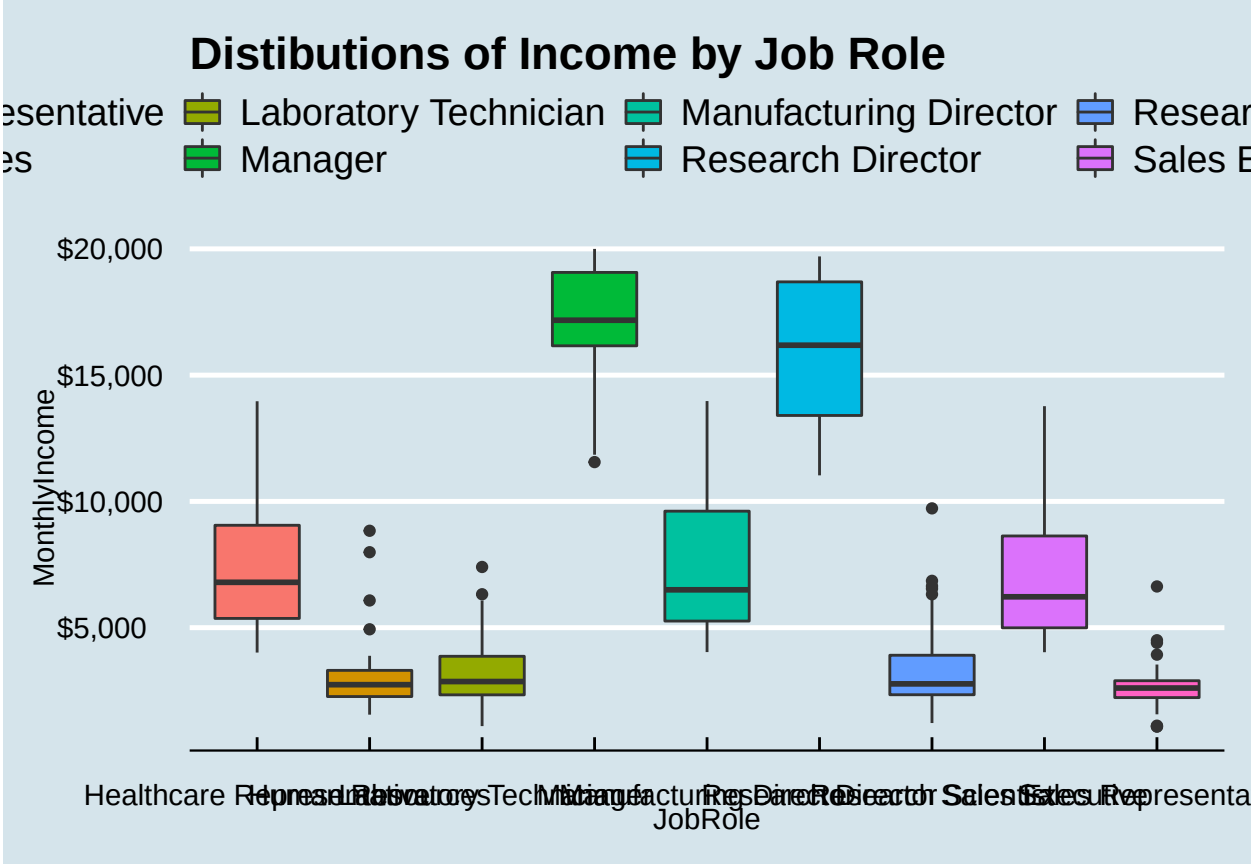



```
#JobSatisfaction in relation to Attrition
data %>% ggplot(aes(x=JobSatisfaction, fill=Attrition)) +
  geom_bar() +
  scale_fill_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  ggtitle("Attrition Rates by Job Satisfaction") +
  scale_y_continuous(labels = scales::percent) +
  labs(y="Attrition (%)")  +
  theme_economist()
```




```
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

##Looking for relationships, BOXPLOTS
data %>% ggplot(aes(x=JobRole, y=MonthlyIncome, fill=JobRole)) +
  geom_boxplot() +
  ggtitle("Distibutions of Income by Job Role") +
  theme(axis.ticks.x = element_blank(),
        axis.text.x = element_blank()) +
  scale_y_continuous(labels = scales::comma)+
  scale_y_continuous(labels=scales::dollar_format()) +
  theme_economist()
```

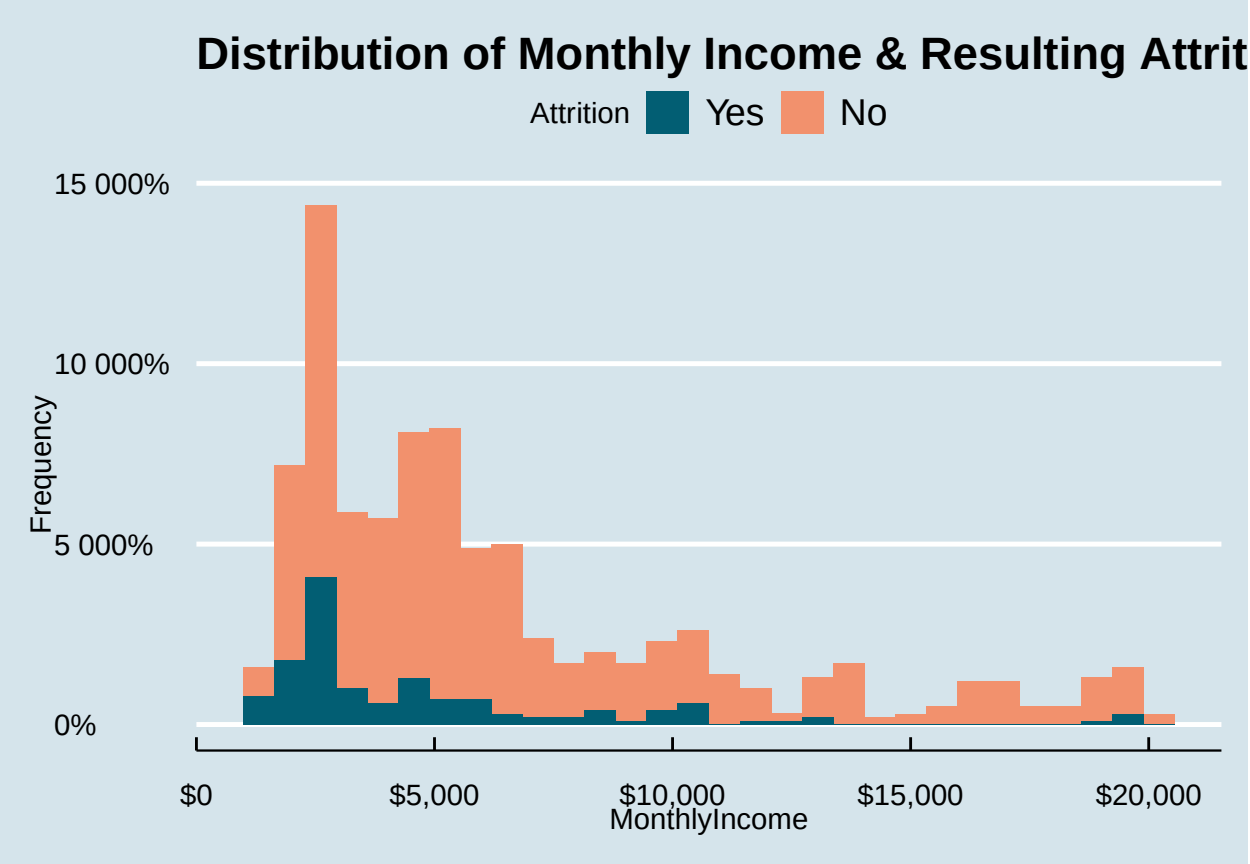
```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```




```
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

##Looking for relationships, HISTOGRAMS

#Distribution of MonthlyIncome, by attrition - Histogram
data %>% ggplot(aes(x=MonthlyIncome, fill=Attrition)) +
  geom_histogram() +
  ggtitle("Distribution of Monthly Income & Resulting Attrition") +
  scale_fill_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  scale_y_continuous(labels = scales::percent) +
  scale_x_continuous(labels = scales::comma)+
  scale_x_continuous(labels=scales::dollar_format()) +
  labs(y="Frequency") +
  theme_economist()
```

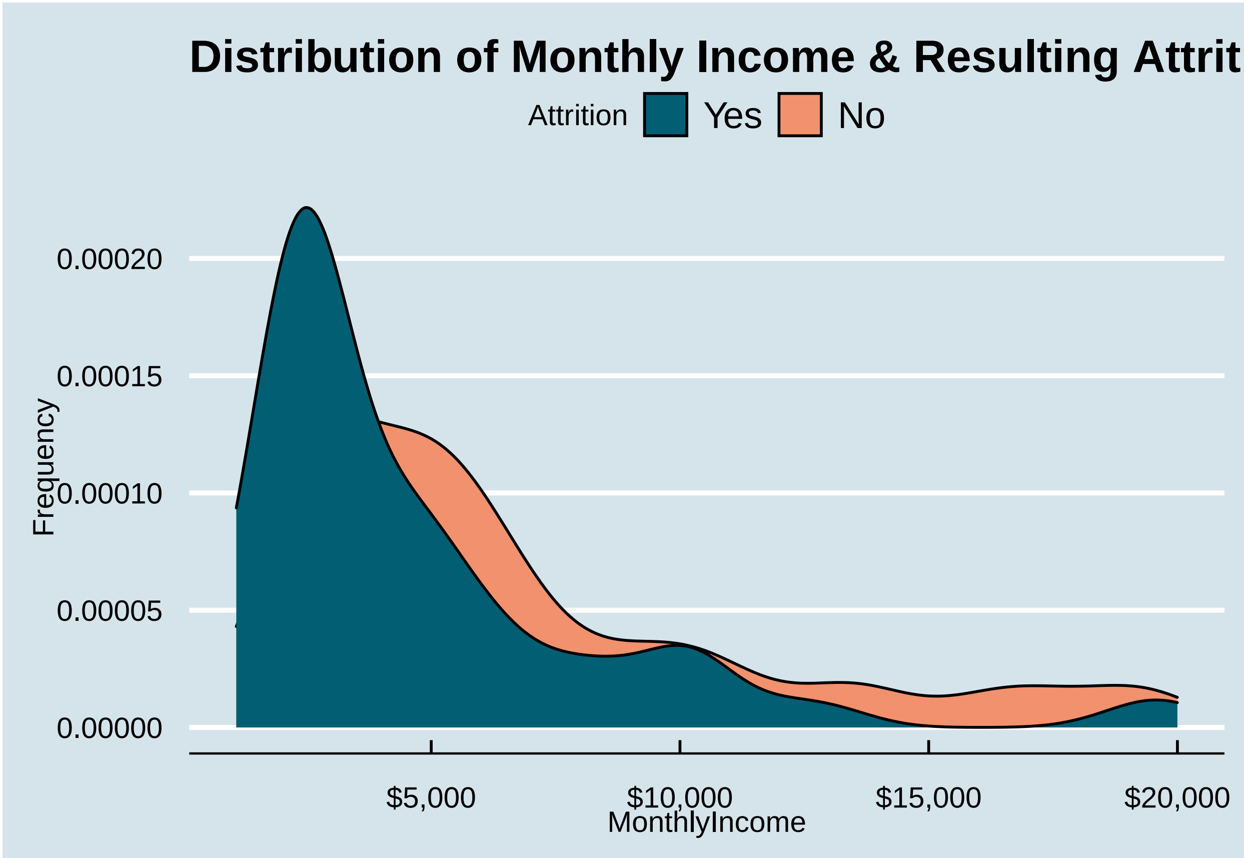
```
## Scale for 'x' is already present. Adding another scale for 'x', which will
## replace the existing scale.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#Distribution of MonthlyIncome, by attrition - Densiry plot
data %>% ggplot(aes(x=MonthlyIncome, fill=Attrition)) +
  geom_density() +
  ggtitle("Distribution of Monthly Income & Resulting Attrition") +
  scale_fill_manual(values = c("Yes" = "#025e73", "No"="#f2916d")) +
  scale_x_continuous(labels = scales::comma)+
  scale_x_continuous(labels=scales::dollar_format()) +
  labs(y="Frequency") +
  theme_economist()
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which will
## replace the existing scale.
```


Distribution of Monthly Income & Resulting Attriti
Attrition
Yes
No
0.00020
0.00015
0.00010
0.00005
0.00000
Frequency
$5,000
$10,000
$15,000
$20,000
MonthlyIncome

```
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # 
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # 

#- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```