Insurify (Take Home Data)

Packages

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:magrittr':
##
##     extract
```

```
library(DAAG)
```

```
## Loading required package: lattice
```

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:dplyr':
##
##     intersect, setdiff, union
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa
```

Reading Data

```
Insurance = read.csv('insurance.csv')
Conversion_rates=read.csv('conversion_rates.csv')
```

Section I: Looking at Demographics + Medical Charges (PART 1)

1.Read in Data and report summary statistics (mean + std / frequency) for age, sex, bmi, children,#smoker, and charges) by region.

Continuous

```
Insurance %>%
    group_by(region) %>%
      summarise(
        count = n(),
        min = min(charges),
        median = median(charges),
        mean = mean(charges),
        std = sd(charges),
        max = max(charges),
        IQR = IQR(charges)
      ) %>%
    arrange(desc(median))
```

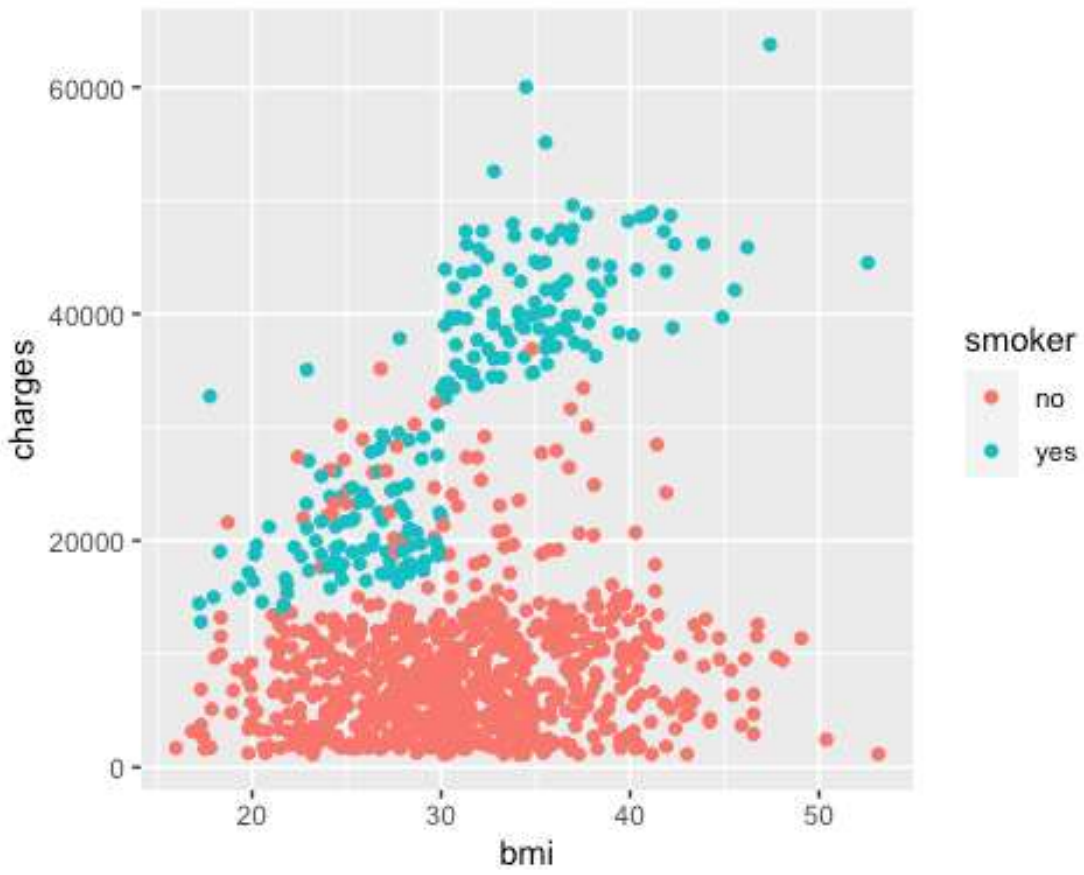
```
## # A tibble: 4 x 8
##   region    count   min median   mean    std    max    IQR
##   <fct>     <int> <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 northeast   225 1695.  9876. 13388. 11126. 48549. 11648.
## 2 southeast   276 1122.  9649. 14953. 13934. 63770. 15044.
## 3 southwest   255 1242.  9145. 12531. 11592. 52591.  9011.
## 4 northwest   248 1621.  8883. 12610. 11329. 60021. 10836.
```

Categorical

```
#Insurance %>% group_by(region) %>% summarise(smoker) %>% table()
```

2.How would you characterize this population? Use figures/tables to support you answer

```
Insurance %>%
    select(smoker, bmi, charges) %>%
    ggplot(aes(color = smoker)) +
    geom_point(mapping = aes(x = bmi, y = charges))
```



Conclusion: People with BMI > 30 and smokes pay high charges

3.In this sample, is female age different from male age?

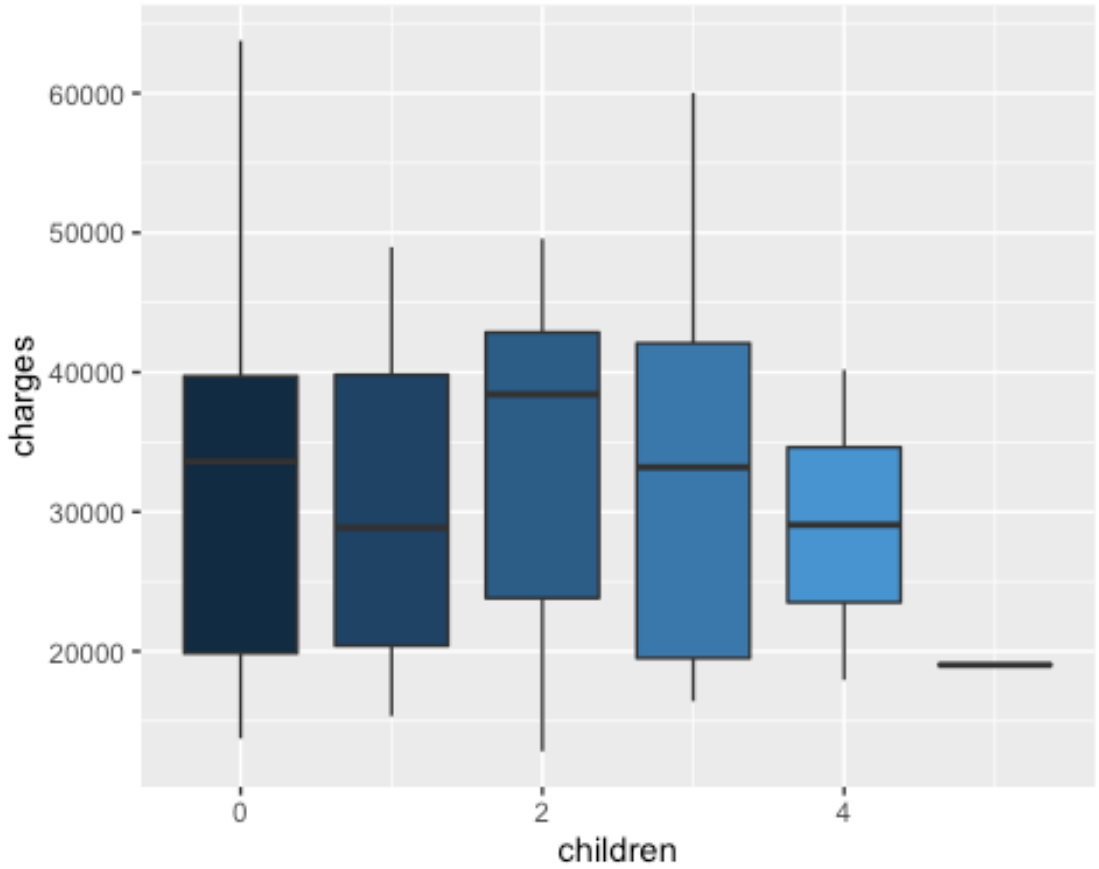
```
Insurance %>%
    group_by(sex) %>%
      summarise(
        count = n(),
        min = min(age),
        median = median(age),
        mean = mean(age),
        median = sd(age),
        max = max(age),
        IQR = IQR(age)
      ) %>%
    arrange(desc(median))
```

```
## # A tibble: 2 x 7
##   sex    count   min median  mean   max   IQR
##   <fct>  <int> <int>  <dbl> <dbl> <int> <dbl>
## 1 male     508    18   14.0  38.6    64  24
## 2 female   496    18   13.9  40.0    64  24.2
```

Conclusion:There is no difference between men and women ages which can been seen from the above statistics.

4.Is there a difference in smoking rates between those who have kids and those who do not?

```
Insurance %>%
    select(children, charges, smoker) %>%
    filter(smoker=="yes")%>%
    ggplot(aes(x = children, y = charges, group = children)) +
    geom_boxplot(outlier.alpha = 0.5, aes(fill = children)) +
    theme(legend.position = "none")
```

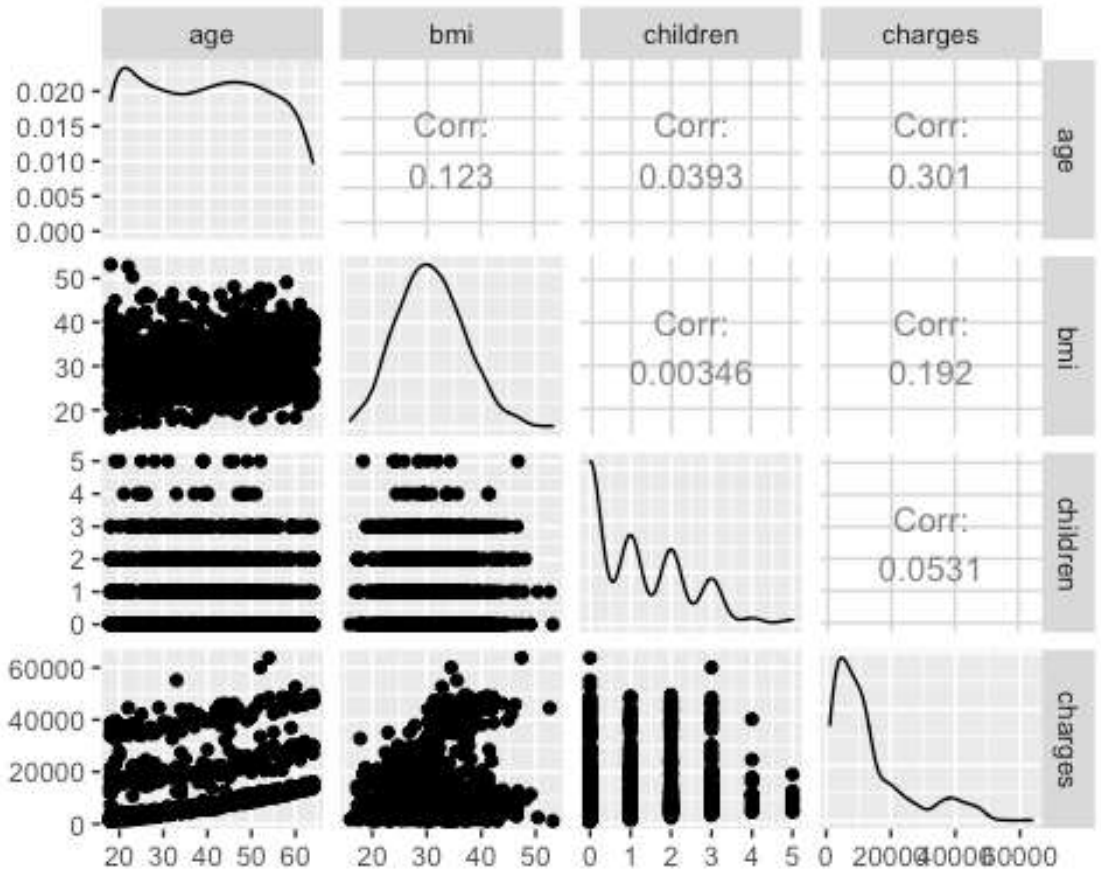


Conclusion: From the above box plot, we conclude that children does not affect charge significantly

5.Are there any instances of high collinearity in this data-set?

```
cor(Insurance$age,Insurance$bmi)

## [1] 0.1225918

cor(Insurance$age,Insurance$children)

## [1] 0.03934207

cor(Insurance$bmi,Insurance$children)
```

```
## [1] 0.003464231

X<-Insurance[,c(1,3,4,7)]
library(GGally)
ggpairs(X)
```



Conclusion: There are no instances of high collinarity in the dataset

6.A coworker wants to know whether:- being male affects medical cost, being a smoker affects medical cost, what is the effect of each additional year on medical cost

```
model<-lm(charges ~ sex + smoker + age, data = Insurance)
summary(model)

##
## Call:
## lm(formula = charges ~ sex + smoker + age, data = Insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15982.7  -2097.8  -1285.1    -80.4  27722.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2886.59     652.78  -4.422 1.08e-05 ***
```

```
## sexmale       289.02     404.55   0.714    0.475
## smokeryes   23633.92     494.34  47.809  < 2e-16 ***
## age           283.34      14.44  19.623  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6379 on 1000 degrees of freedom
## Multiple R-squared:  0.725,  Adjusted R-squared:  0.7242
## F-statistic: 878.8 on 3 and 1000 DF,  p-value: < 2.2e-16
```

Conclusion:From the above Regression model, we conclude that

1)Being male doesn't affect the model,

2)Being smoker does affect the model and

3) The medical cost increases as the age increases.

On the basis of lev of significance,it is observed that age and smoker are statistically significant as compared to sex as the p-values are lesser than level of significance (0.05)

7.Your boss comes to you and says we want to limit patients that may cost more than 50K. You don't need to write code to do this but outline how you could create a model that would take a new patient's characteristics and output the probability that their medical charges would be over 50K.

I would create a new column say "Charges_flag" where the charges >50000 will be 1 and <50000 will be 0. Considering "Charges_flag" as the target column and the other characteristics as the input columns, calculate the probability using any binary classification algorithm like SVM, Binomial Logistic Regression.

The evaluate the effectiveness of the model I would calculate the confusion matrix (accuracy, precision, recall, F1 score)

On basis of the output of the model, we will define the value of X, beyond which the the patient would cost>50,000.

Section II: Checking Conversion Rates (PART 2)

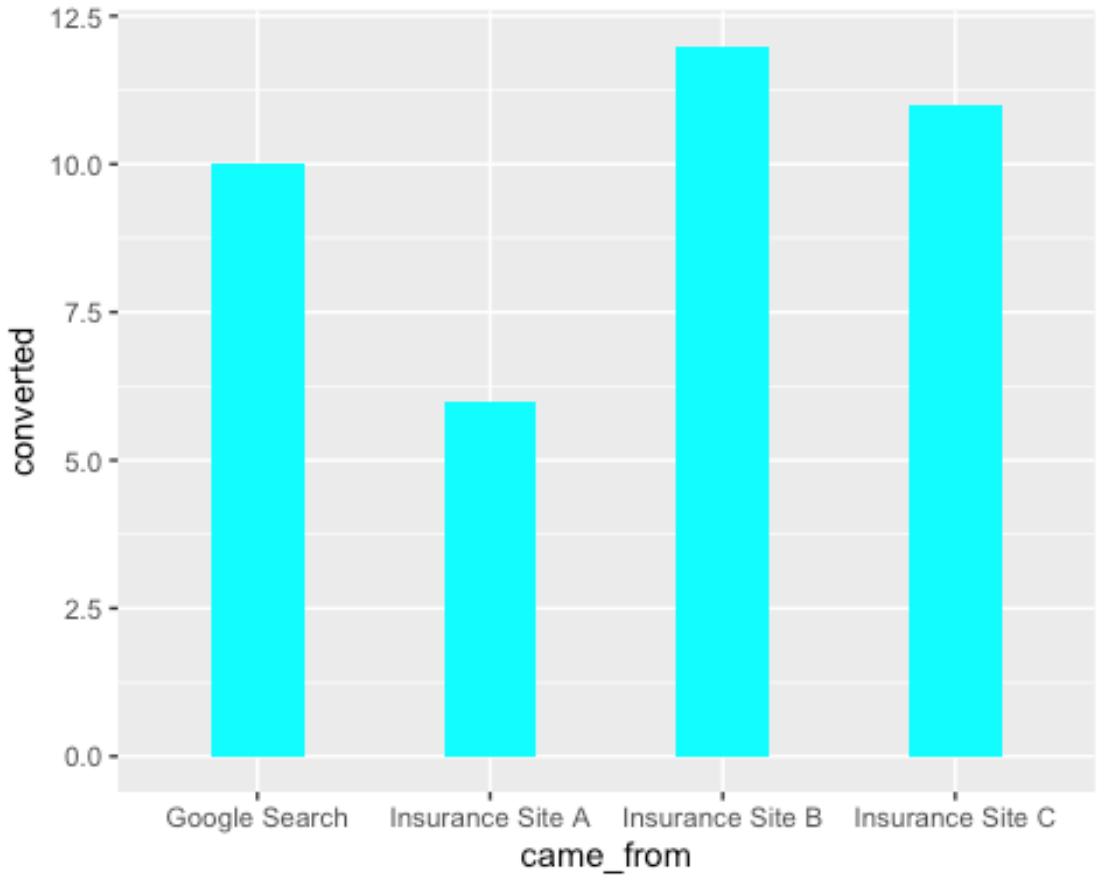
Conversion Rates

```
Conversion_rates$converted = ifelse(Conversion_rates$has_insurance == 0 &
Conversion_rates$reached_end == 1,1,0 )

Conversion_rates$date <- as.Date(Conversion_rates$date, format = "%Y-%m-%d")
```

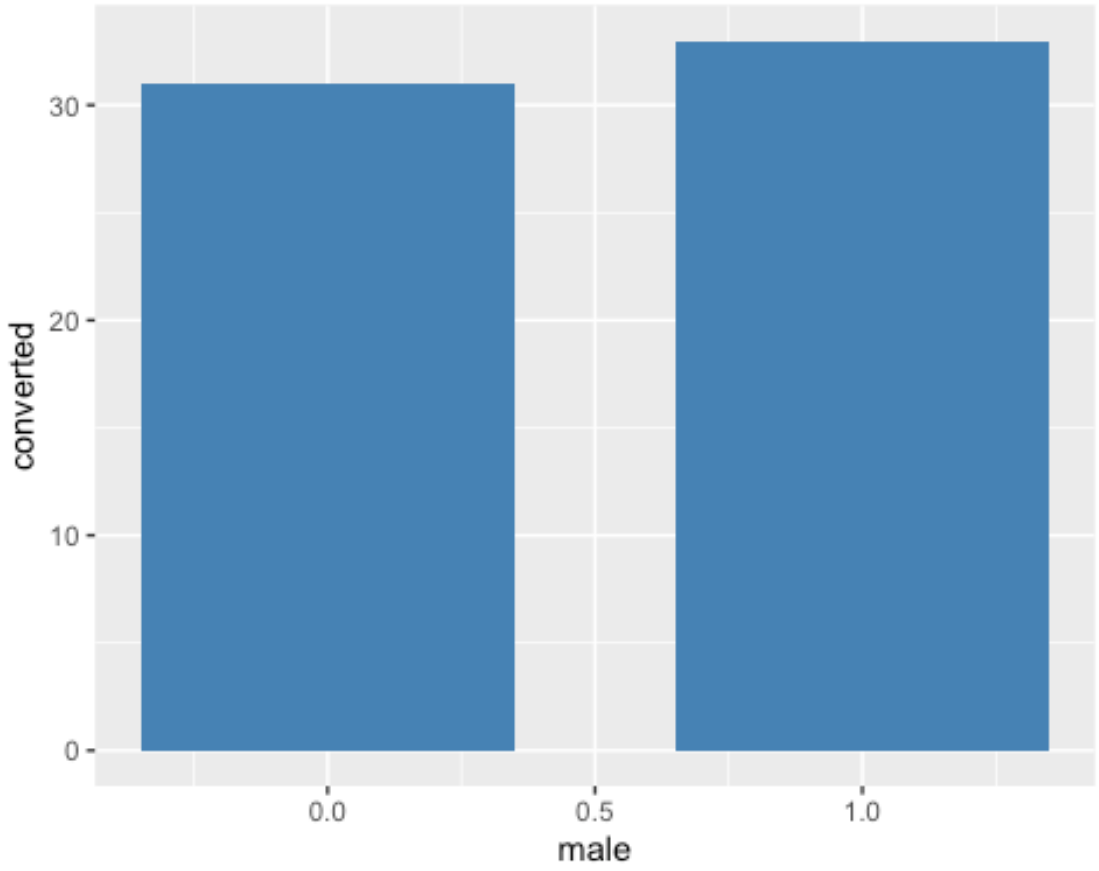
```r
before <- sum(Conversion_rates[which(Conversion_rates$date < ymd('2018-09-05')), 7])
before<- before/nrow(Conversion_rates)
after <- sum(Conversion_rates[which(Conversion_rates$date >= as.Date('2018-09-05')), 7])
after <- after/nrow(Conversion_rates)
```

Conclusion: The conversion rate from 5th onwards after the new feature launch is 13.9% whereas the rate before the feature launch is 8.9% on the overall dataset.

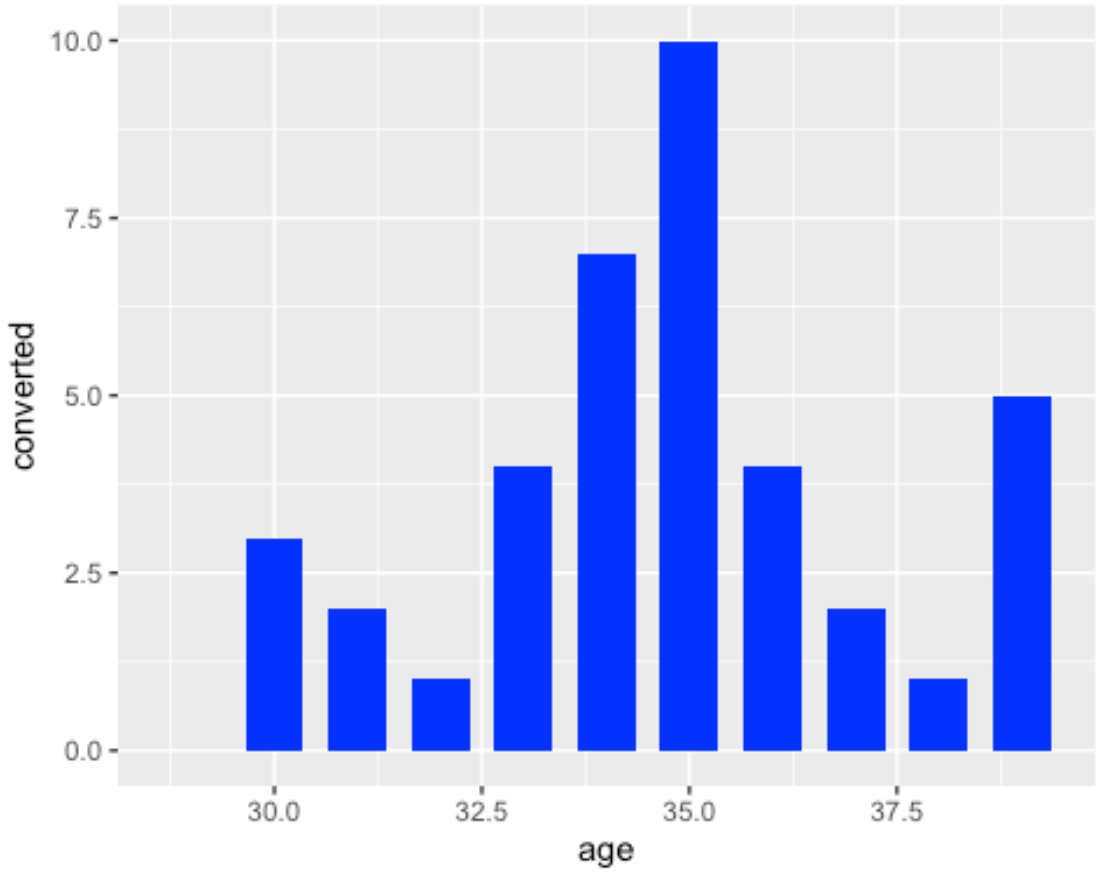
```r
altered<-Conversion_rates%>%

  filter(date >= as.Date("2018-09-05"))

ggplot(altered, aes(x=came_from, y=converted)) +
  geom_bar(stat = "identity",width=0.4,fill="cyan")
```




```r
ggplot(Conversion_rates, aes(x=male, y=converted)) +
  geom_bar(stat = "identity",width=0.7,fill="steelblue")
```




```
ggplot(altered, aes(x=age, y=converted)) +
  geom_bar(stat = "identity",width=0.7,fill="blue")
```



Recommendation : Company can target customers of age 34 and 35 years as there is a high probabilty of conversion.There is no difference observed between gender.Another recommention is to check what different features are used by Insurance Site A, as least customers are converted who visited from Insurance Site A.

Section III - Visualizing Data (PART 3)

The single figure that would help the executive team to grow the business is as follows:

(Age is binned with an interval of 5 years for customers w/o policy bought. Statistical analysis was then performed on the Policy amount and is presented in the figure below. Error bars denote the standard deviation)

(Figure is generated using Excel, hence attaching the .xlsx file as well)

Insights:

1) Highest sale is made by the age group 40-45 with the min policy value of $300 and average value from all 3 groups is $447.

2) Even though group C had the lowest intent, overall customers from group C bought the policy with an equivalent amount of group A, who had the highest intent.