Assignment 1

Travis Rutledge, Dylan Cheung, Dulitha Perera, Chris Liolios

2024-09-24

Task 1

To gain a better understanding of our customers' incomes, we can look at descriptive statistics, a histogram, and some fitted models of their data.

Descriptive Statistics and Distribution of Customers' Income

First, let's look at some descriptive statistics of customers' income.

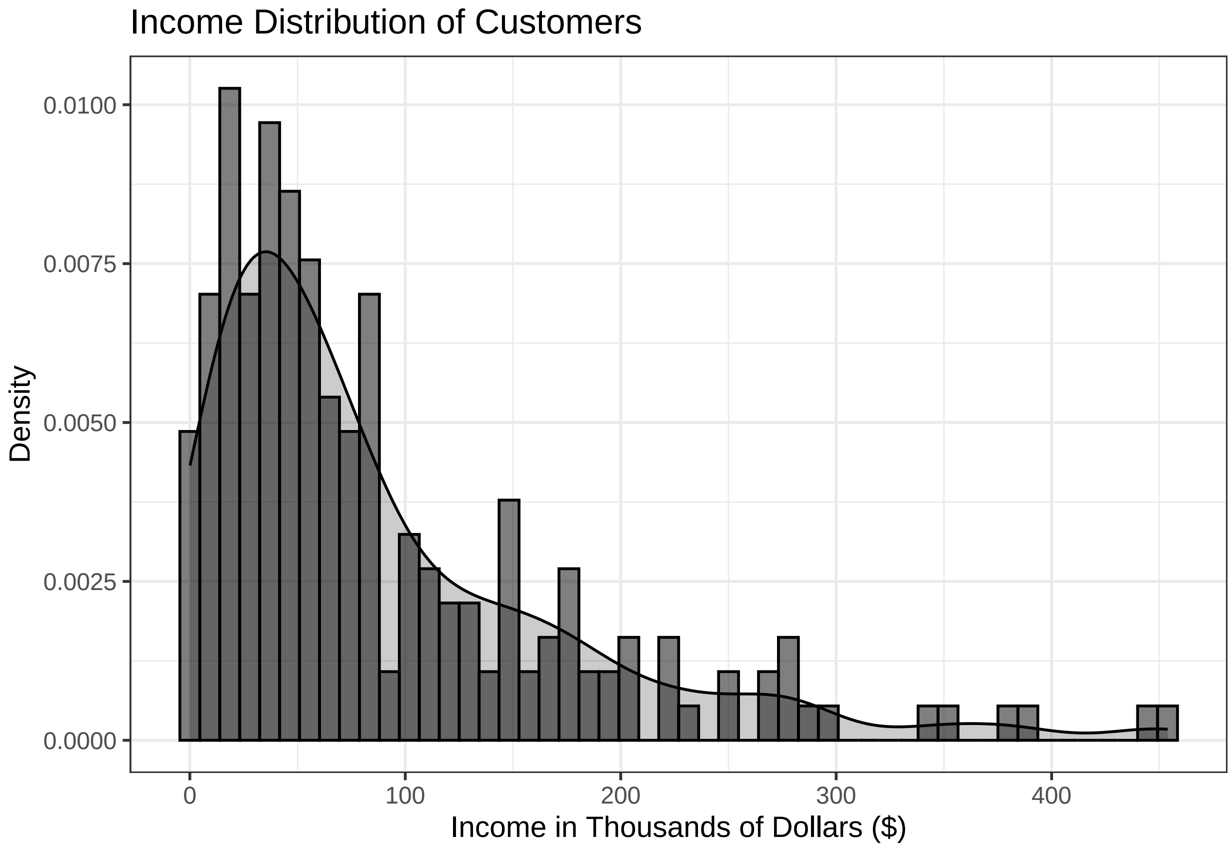
```
bankdata %>%
summarise(n = n(), mean = mean(income), median = median(income),
SD = sd(income), IQR = IQR(income)) %>% kable(digits = 1) %>%
kable_styling(latex_options = "hold_position")
```

n	mean	median	SD	IQR
200	89.2	58.7	87.6	94

Of the 200 customers sampled, the average income is $89k and the median income is $58k, which suggests that there are some very high income customers who are bringing the average higher than the median. The standard deviation (SD) tells us how spread out the income levels are around the mean of $89k. For our customers, the income deviates from the mean by $87.6k, on average. The interquartile range (IQR) tells us how spread the income is for the middle 50% of the data. In this dataset, the middle 50% of customers fall within a range of $94k.

Next, let's look at a histogram visualising the income of customers at all education levels. This histogram shows that the distribution of customers' income is right-skewed and that most customers have incomes between $0k-$100k. As the level of income increases, the number of customers' who earn those higher incomes decrease.

```
histogram_dataplot <- function(x, bins = 50, fill = "grey", colour = "black") {
  ggplot(tibble(x = x), aes(x = x, y = after_stat(density))) +
    geom_histogram(bins = bins, colour = colour, fill = colour, alpha = 0.5) +
    geom_density(colour = "black", fill = colour, alpha = 0.2) +
    labs(title = "Income Distribution of Customers",
         x = "Income in Thousands of Dollars ($)", y = "Density") +
    theme_bw()
}
income_histogram <- histogram_dataplot(bankdata$income)
income_histogram
```



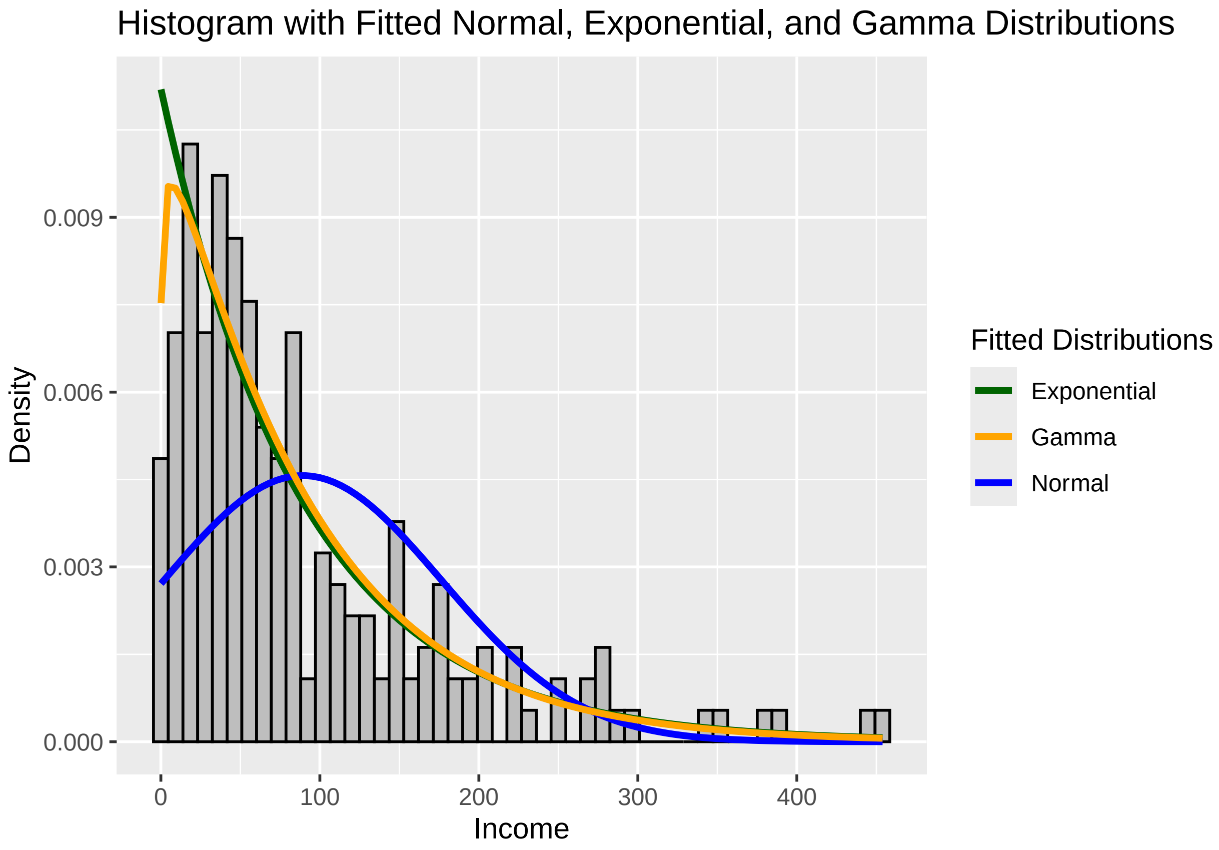
Fitted Models - Normal, Exponential, and Gamma

To further analyze the data, three models were fitted using Maximum Likelihood Estimation (MLE) to determine which one best describes the income distribution of our customers. Fitting a distribution model allows us to make predictions about the population based on our sample of 200 customers.

For the normal distribution, MLE estimated the mean as $89.2K and the standard deviation as $87.4K. For the exponential distribution, MLE estimated the rate parameter as .0112, which means that income decreases at a rate of 1.12% per unit of income. Lastly, for the gamma distribution, MLE estimated the shape parameter as 1.075 and rate parameter as .012, suggesting that income increases at the start of the distribution and then gradually decreases.

The histogram below shows the income distribution overlaid with the three fitted distribution curves overlaid. Among these, the gamma distribution (orange line) appears to be the best fit, as it captures the sharp rise in the number of customers with incomes between $0k and $60k, followed by a gradual decline as income levels increase.

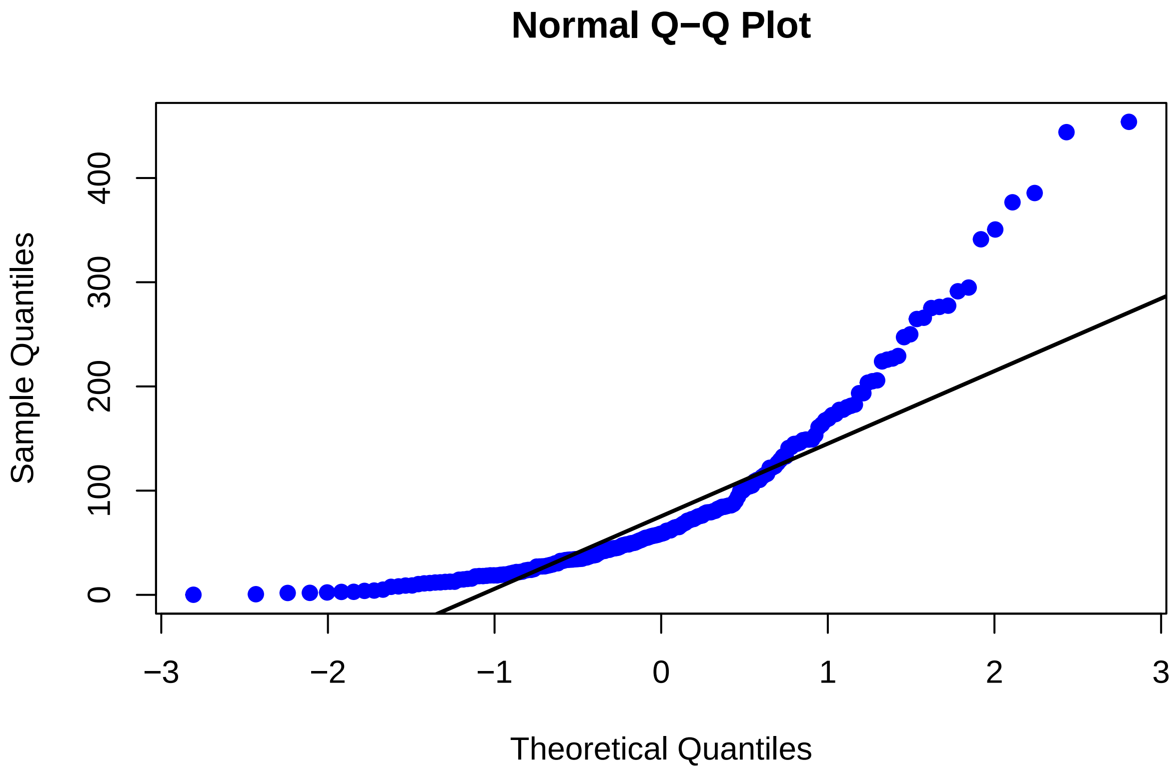
```
normal_fit <- fitdistr(bankdata$income, "normal")
exp_fit <- fitdistr(bankdata$income, "exponential")
gamma_fit <- fitdistr(bankdata$income, "gamma")

ggplot(bankdata, aes(x = income)) +
  geom_histogram(aes(y = ..density..), bins = 50, fill = "grey", color = "black") +
  stat_function(fun = dexp, args = list(rate = exp_fit$estimate[1]),
                aes(colour = "Exponential"), size = 1.2) +
  stat_function(fun = dnorm, args = list(mean = normal_fit$estimate[1],
                                         sd = normal_fit$estimate[2]),
                aes(colour = "Normal"), size = 1.2) +
  stat_function(fun = dgamma, args = list(shape = gamma_fit$estimate[1],
                                          rate = gamma_fit$estimate[2]),
                aes(colour = "Gamma"), size = 1.2) +
  labs(title = "Histogram with Fitted Normal, Exponential, and Gamma Distributions",
       x = "Income", y = "Density", color = "Fitted Distributions") +
  scale_color_manual(values = c("Normal" = "Blue",
                                "Exponential" = "Darkgreen",
                                "Gamma" = "Orange"))
```



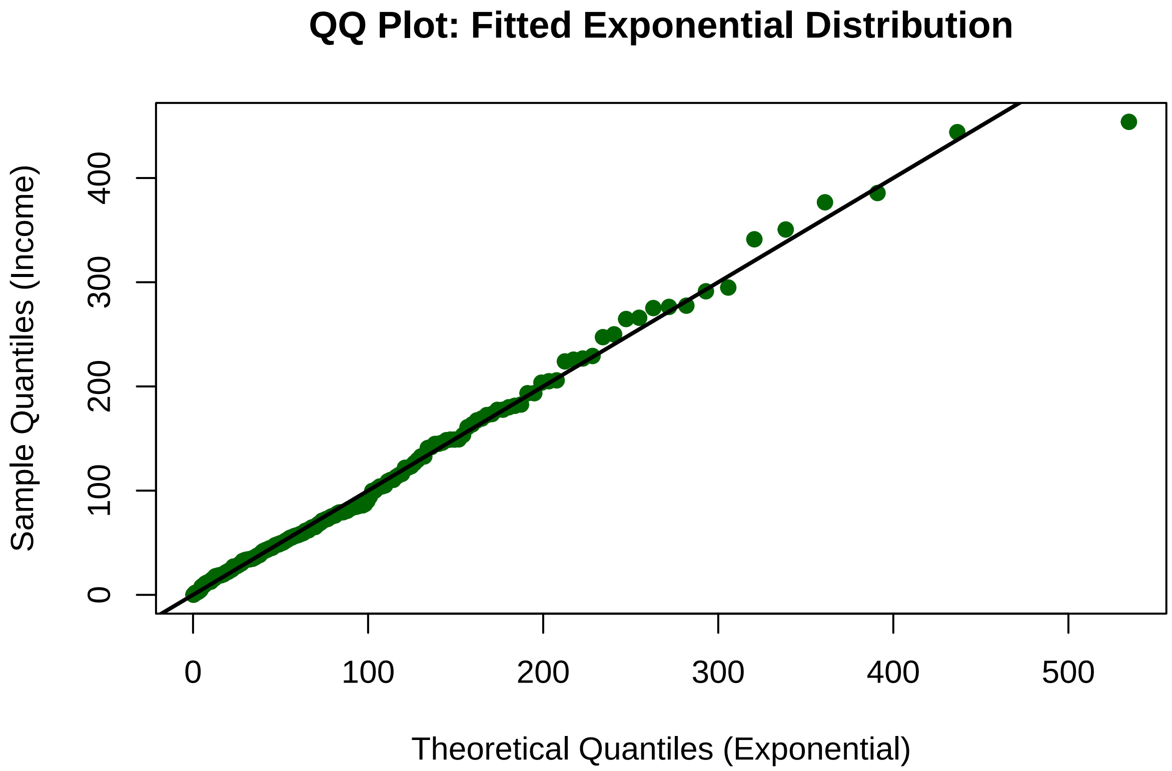
QQ Plots

To further evaluate which fitted model best describes the customer income data, we can use Quantile-Quantile (QQ) plots. A QQ plot compares the quantiles of the customer income data with the theoretical quantiles from the normal, exponential, or gamma distributions. If the points on the QQ plot follow a straight line, it suggests that the model fits the data well. Below are the QQ plots for the normal, exponential, and gamma distributions. The plot that follows the straightest line is for the gamma distribution, providing additional evidence that it is the best-fitting model for the customer income data.

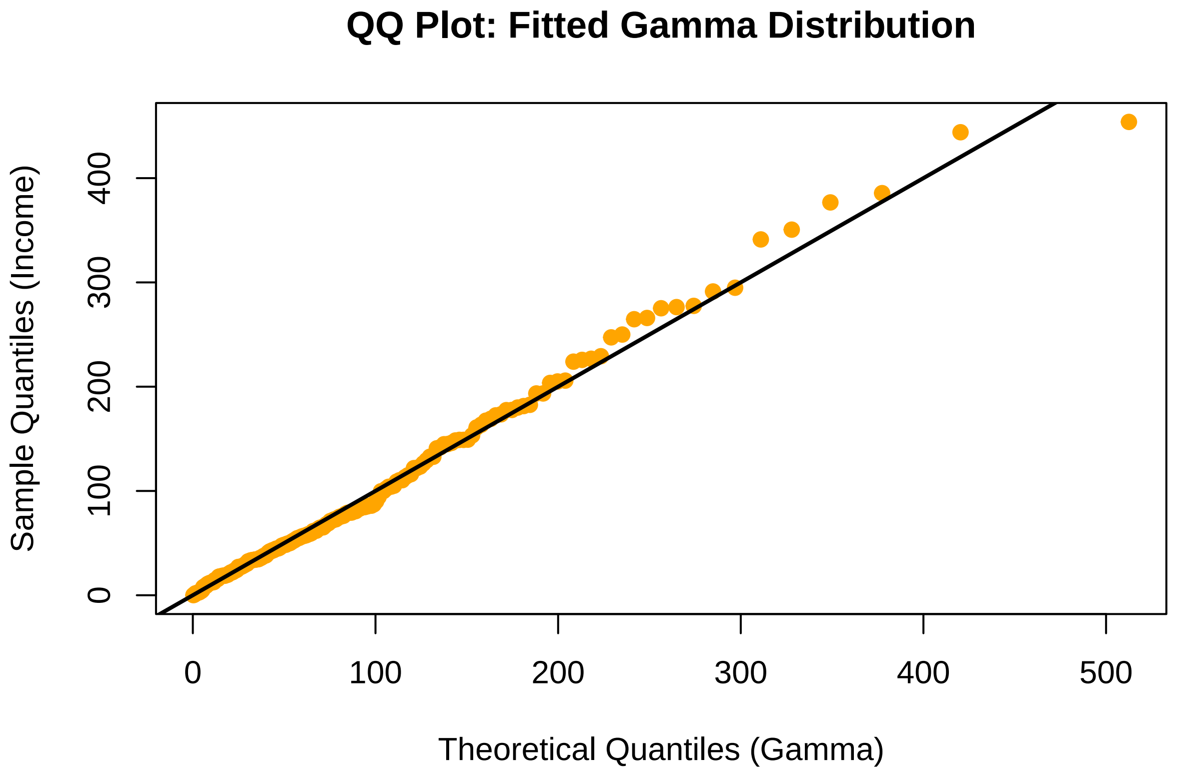
```
#normal QQ
qqnorm(bankdata$income, col = "blue", pch = 19)
qqline(bankdata$income, color = "black", lwd = 2)
```



```r
#exponential QQ
qqplot(qexp(ppoints(length(bankdata$income)), rate = exp_fit$estimate[1]),
       sort(bankdata$income),
       main = "QQ Plot: Fitted Exponential Distribution",
       xlab = "Theoretical Quantiles (Exponential)",
       ylab = "Sample Quantiles (Income)",
       col = "darkgreen", pch = 19)
abline(0, 1, col = "black", lwd = 2)
```



```r
#gamma QQ
qqplot(qgamma(ppoints(length(bankdata$income)),
        shape = gamma_fit$estimate[1], rate = gamma_fit$estimate[2]),
       sort(bankdata$income),
       main = "QQ Plot: Fitted Gamma Distribution",
       xlab = "Theoretical Quantiles (Gamma)",
       ylab = "Sample Quantiles (Income)",
       col = "orange", pch = 19)
abline(0, 1, col = "black", lwd = 2)
```

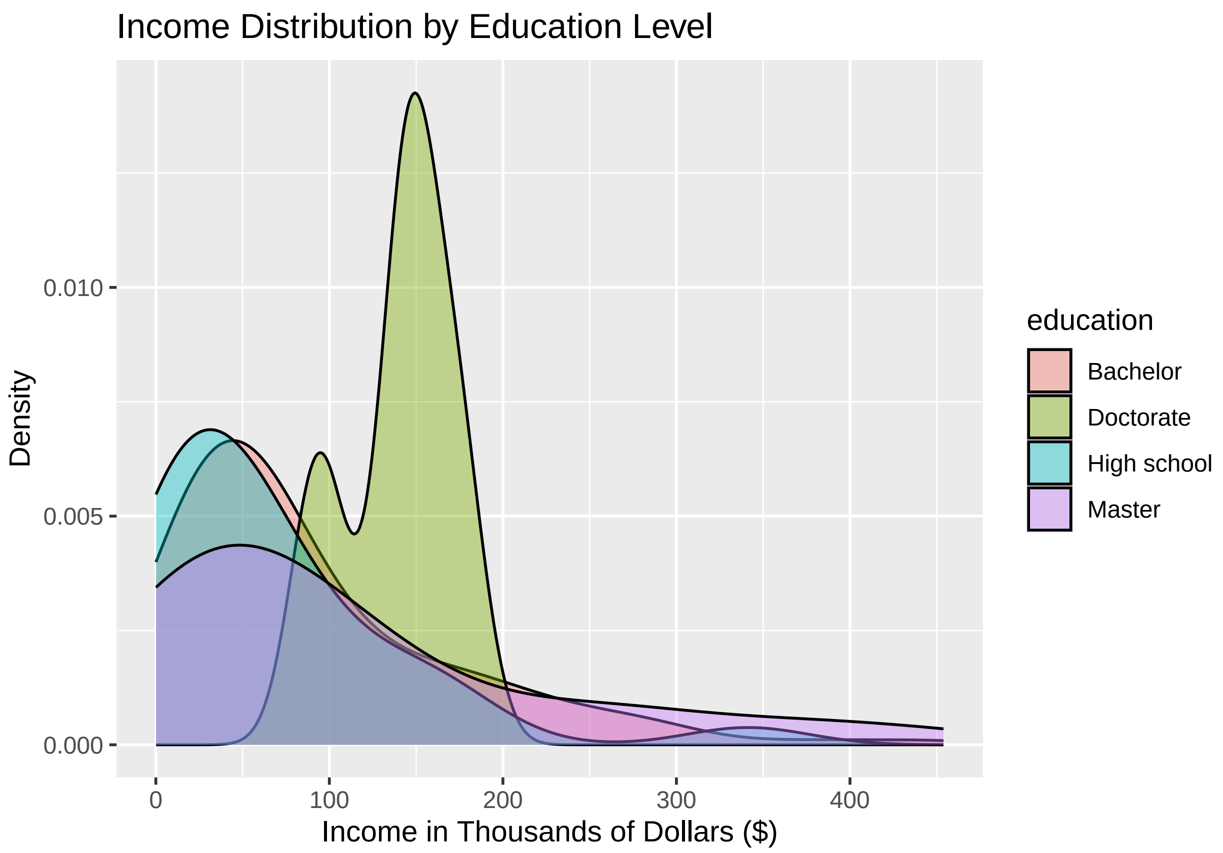


Task 4

Broken Out by Education Level

When we break out this distribution by education level, we can see a more informed story. Generally speaking, as the level of education increases, so does the level of income. Customers who have doctorates have the highest income, on average. The following is a density plot showing customer income broken out by education level:

```
densityplot <- ggplot(data=bankdata, aes(x=income, group=education, fill=education)) +
    geom_density(adjust=1.5, alpha = 0.4) +
    labs(title = "Income Distribution by Education Level",  # Title
         x = "Income in Thousands of Dollars ($)",
         y = "Density")

densityplot
```



Below is a table providing descriptive statistics of customer income for each education level.

```r
bankdata_wider <- pivot_wider(bankdata, names_from = education, values_from = income)

kable(summary(bankdata_wider))
```

customer	Master	Bachelor	High school	Doctorate
Length:200	Min. : 1.80	Min. : 1.90	Min. : 0.10	Min. : 94.2
Class :character	1st Qu.: 28.88	1st Qu.: 31.95	1st Qu.: 12.45	1st Qu.:132.3
Mode :character	Median : 68.35	Median : 58.70	Median : 45.05	Median :145.6
NA	Mean :109.99	Mean : 87.84	Mean : 65.06	Mean :139.7
NA	3rd Qu.:142.28	3rd Qu.:122.28	3rd Qu.: 81.50	3rd Qu.:152.9
NA	Max. :453.90	Max. :444.00	Max. :341.20	Max. :173.5
NA	NA's :166	NA's :68	NA's :170	NA's :196

As wee see in the density plot above, the average and median incomes increase as the level of education increases. Although customers with doctorates earn the highest average wages, they do not have the highest maximum income in the dataset. The maximum income in the dataset is $453.90k, and that comes from a customer with a Master's degree. Customer's with Master's degrees have the highest variability in income compared to ther education levels.