Lab I - Statistics and Data Analysis II

Zhen Liu

November 11, 2024

This lab is done by Zhen Liu

Task 1

```
# Load the dataset. Assign data to objects.
hsbc_basic <- read.csv("hsbc-basic.csv", header = TRUE)
hsbc_health <- read.table("hsbc-health.txt", header = TRUE)
```

```
# number of rows and columns
cat("hsbc_basic: Rows-", nrow(hsbc_basic), "Columns -", ncol(hsbc_basic), "\n")
```

```
## hsbc_basic: Rows- 2000 Columns - 3
```

```
cat("hsbc_health: Rows-", nrow(hsbc_health), "Columns -", ncol(hsbc_health), "\n")
```

```
## hsbc_health: Rows- 1500 Columns - 4
```

```
# data types for hsbc_basic
print("hsbc_basic - Variable Types:\n")
```

```
## [1] "hsbc_basic - Variable Types:\n"
```

```
str(hsbc_basic)
```

```
## 'data.frame':    2000 obs. of  3 variables:
##  $ id4   : int  3303 7443 1297 4775 1906 1149 2749 2391 3773 4944 ...
##  $ sex   : chr  "Boy" "Girl" "Girl" "Boy" ...
##  $ AGECAT: int  13 15 11 13 11 11 13 11 13 13 ...
```

```
# data types for hsbc_health
print("hsbc_health - Variable Types:\n")
```

```
## [1] "hsbc_health - Variable Types:\n"
```

```r
str(hsbc_health)
```

```
## 'data.frame':    1500 obs. of  4 variables:
##  $ id4         : int  2 27 35 37 38 39 40 41 49 60 ...
##  $ bully_dummy : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ health_index: int  8 8 8 6 6 7 7 6 10 6 ...
##  $ lifesat     : num  6.9 9.84 6.81 9.91 9.62 ...
```

```r
# Merge datasets on id4
hsbc <- merge(hsbc_basic, hsbc_health, by = "id4")
# Display number of rows and columns in the merged data
cat("hsbc - Rows:", nrow(hsbc), "Columns:", ncol(hsbc), "\n")
```

```
## hsbc - Rows: 1500 Columns: 6
```

Since we used an inner join, the number of rows in hsbc is determined by matching records in both hsbc_basic and hsbc_health

```r
# Find columns with missing values in the merged dataset
missing_values <- sapply(hsbc, function(x) sum(is.na(x)))
missing_columns <- names(missing_values[missing_values > 0])

# Display columns with missing values and the number of missing entries
print("Columns that misses values - ")
```

```
## [1] "Columns that misses values - "
```

```r
print(missing_values[missing_columns])
```

```
## lifesat
##      10
```

```r
# Average life satisfaction
average_lifesat <- mean(hsbc$lifesat, na.rm = TRUE)
cat("Average Life Satisfaction -", average_lifesat, "\n")
```

```
## Average Life Satisfaction - 7.344637
```

```r
# Observations by age category
age_counts <- table(hsbc$AGECAT)
print(age_counts)
```

```
##
##  11  13  15
## 474 447 579
```

```r
# Age category with most observations
most_observed_agecat <- names(which.max(age_counts))
cat("Age category with most observations:", most_observed_agecat, "\n")
```

```
## Age category with most observations: 15
```

15 has the most observations

```
# Bully by age category
bullying_counts <- table(hsbc$AGECAT[hsbc$bully_dummy == 1])
print(bullying_counts)
```

```
##
## 11 13 15
## 77 54 45
```

```
# Age category with most bullied children
most_bullied <- names(which.max(bullying_counts))
cat("highest recorded number category of bullied kids", most_bullied, "\n")
```

```
## highest recorded number category of bullied kids 11
```

11 has the highest recorded number of bullied kids

```
# (a)
low_lifesat_bullied <- nrow(subset(hsbc, bully_dummy == 1 & lifesat < 7))
cat("Bullied kids with lifesat < 7:", low_lifesat_bullied, "\n")
```

```
## Bullied kids with lifesat < 7: 95
```

```
# (b)
high_lifesat_girls <- nrow(subset(hsbc, sex == "Girl" & AGECAT == 13 & lifesat > 8))
cat("Girls in age 13 with lifesat > 8:", high_lifesat_girls, "\n")
```

```
## Girls in age 13 with lifesat > 8: 77
```

95 bullied kids are with a lifesat score lower than 7 77 girls in age-category 13 (AGECAT==13) that have a lifesat score greater than 8

```
hsbc$health_index_binary <- ifelse(hsbc$health_index >= 7, 1, 0)
head(hsbc$health_index_binary)
```

```
## [1] 1 1 1 0 0 1
```

```
# conditional mean of lifesat
lifesat_means <- aggregate(lifesat ~ health_index_binary, data = hsbc, mean)
print(lifesat_means)
```

```
##   health_index_binary  lifesat
## 1                   0 6.264939
## 2                   1 7.817786
```

```r
highest_lifesat_status <- lifesat_means$health_index_binary[which.max(lifesat_means$lifesat)]
cat("highest average life satisfaction for binary of...", highest_lifesat_status, "\n")
```

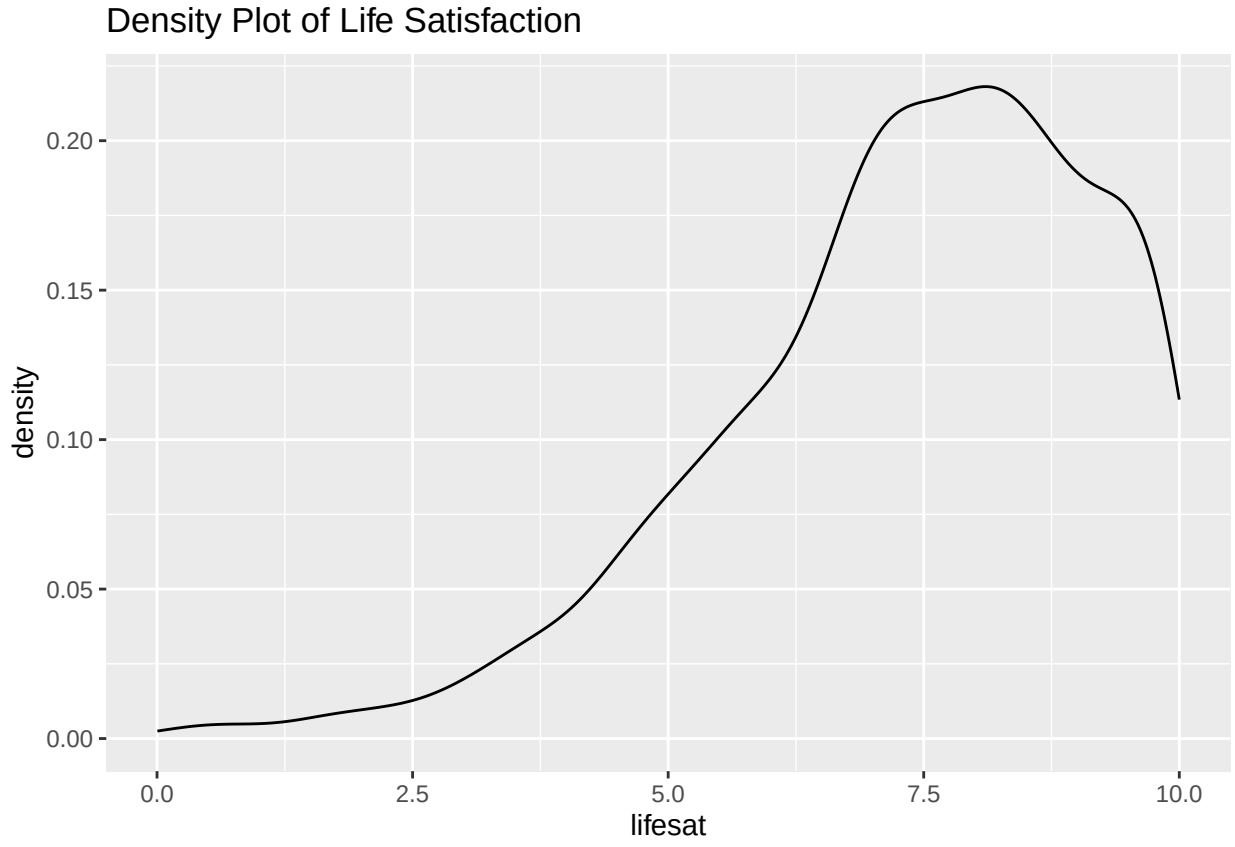
```
## highest average life satisfaction for binary of... 1
```

Conditional means of lifesat by health_index_binary: 0: 6.264939 1: 7.817786

highest average life satisfaction: 1

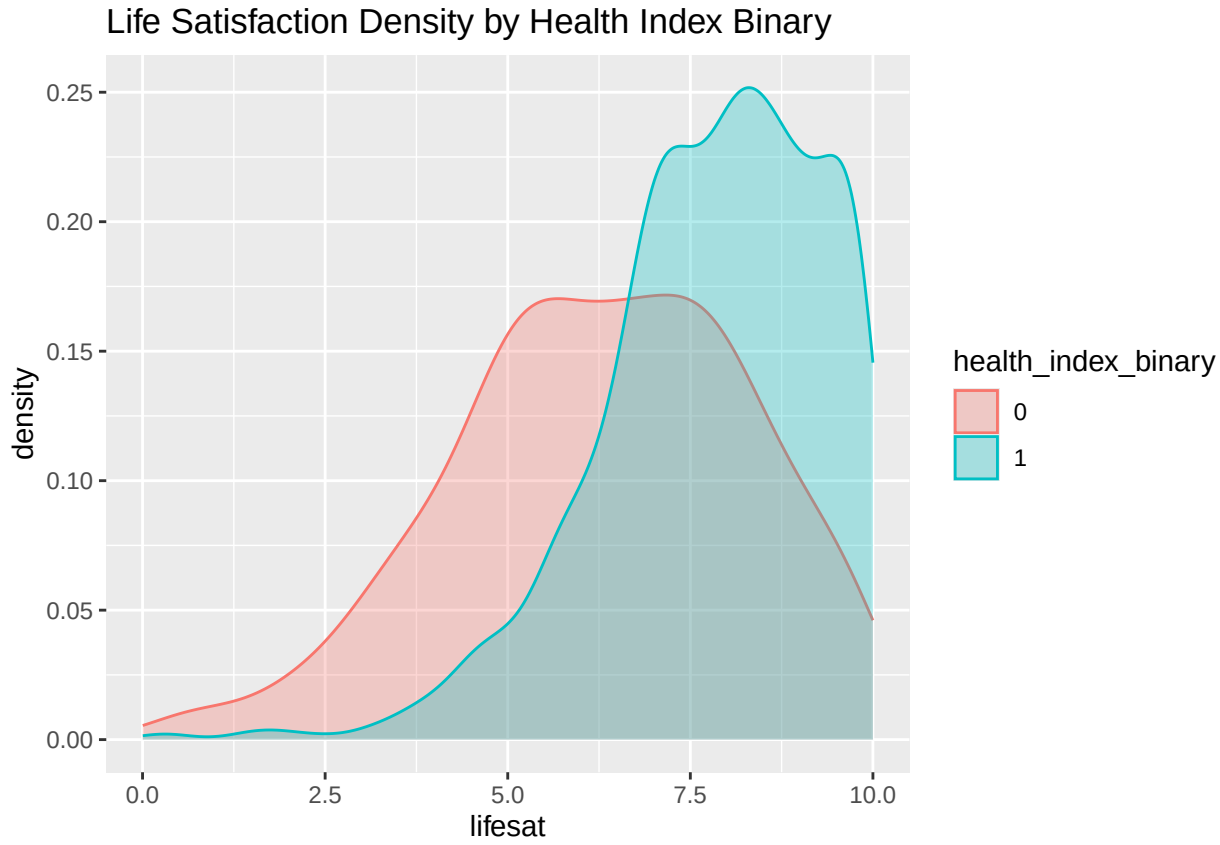
```r
library(ggplot2)

# health_index_binary to Factor
hsbc$health_index_binary <- factor(hsbc$health_index_binary)

# Plot density of lifesat
ggplot(hsbc, aes(x = lifesat)) +
    geom_density() +
    labs(title = "Density Plot of Life Satisfaction")
```

```
## Warning: Removed 10 rows containing non-finite outside the scale range
## (‘stat_density()‘).
```



```r
# Density plot
ggplot(hsbc, aes(x = lifesat, color = health_index_binary, fill = health_index_binary)) +
    geom_density(alpha = 0.3) +
    labs(title = "Life Satisfaction Density by Health Index Binary")
```

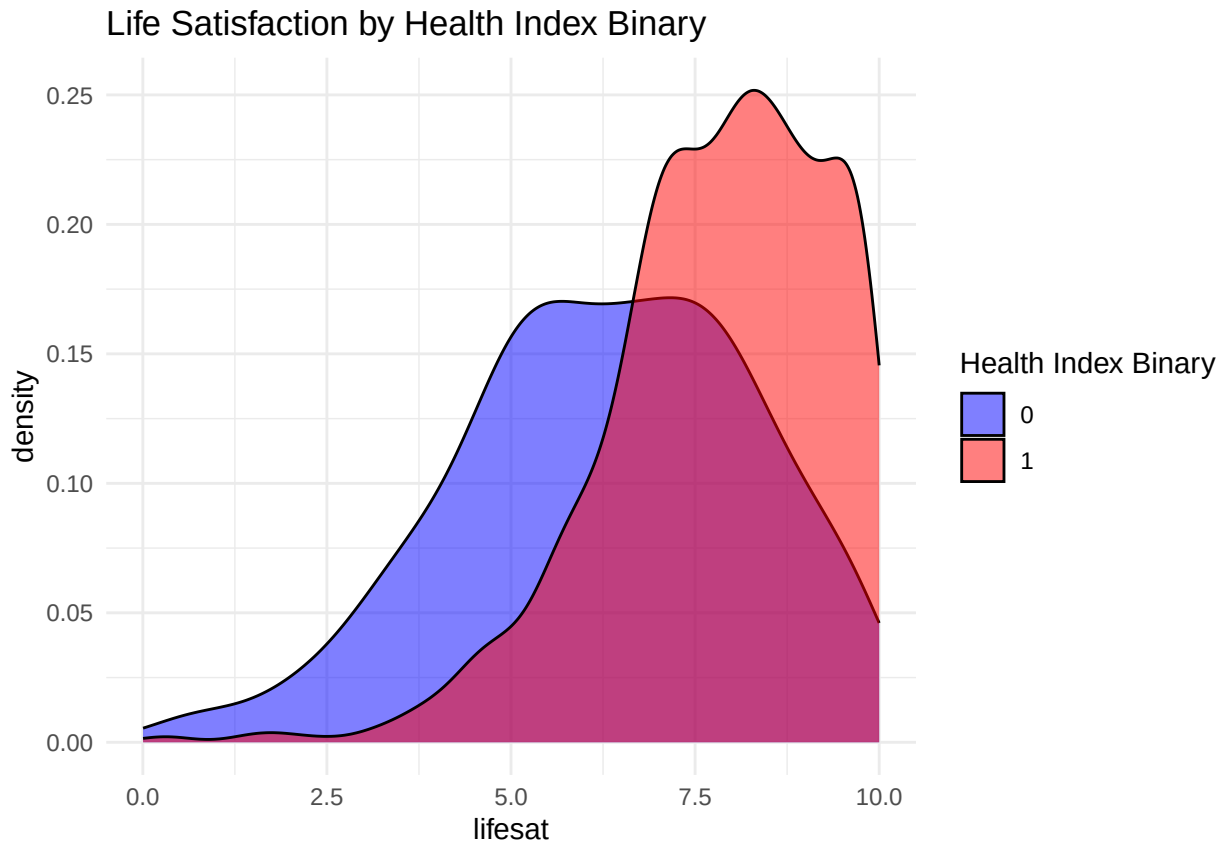
```
## Warning: Removed 10 rows containing non-finite outside the scale range
## (‘stat_density()‘).
```



The distribution of is right-skewed, suggesting that most individuals from the data sample report moderate satisfaction, with just few reporting high satisfaction

```r
# Load ggplot2 library for plotting
library(ggplot2)

# Density plot of life satisfaction, colored by health_index_binary
ggplot(hsbc, aes(x = lifesat, fill = health_index_binary)) +
  geom_density(alpha = 0.5) +
  labs(title = "Life Satisfaction by Health Index Binary",
       x = "lifesat",
       y = "density") +
  scale_fill_manual(values = c("blue", "red"), name = "Health Index Binary") +
  theme_minimal()
```

```
## Warning: Removed 10 rows containing non-finite outside the scale range
## (‘stat_density()‘).
```



This plot shows the distribution of life satisfaction scores for individuals with different health index binary statuses (0 and 1).

```
write.csv(hsbc, "hsbc.csv", row.names = FALSE)
```