Bivariate Kernel Density Estimation Report: Titanic Dataset

Author: Linh Chi Tran
Date: April 11, 2025

1. Data Description and Source

The dataset used in this analysis is the **Titanic training dataset**, obtained from the Kaggle Titanic: Machine Learning from Disaster competition. Link to the data set:
(https://www.kaggle.com/competitions/titanic/data)

It contains detailed information on passengers aboard the Titanic, including these variables:

- survival(survival)
- pclass (ticket class)
- sex(sex)
- age(age in years)
- sibsp(of siblings / spouses aboard the Titanic)
- parch(of parents / children aboard the Titanic)
- ticket(ticket number), fare(passenger fare)
- cabin(cabin number)
- embarked (port of embarkation).

For the purpose of this kernel density analysis, two continuous variables were selected:

Age: Passenger's age in years
Fare: Ticket price in British Pounds
Rows with missing values in either Age or Fare were excluded to ensure a complete-case analysis, resulting in a clean dataset suitable for bivariate density estimation.

2. Plots of the Bivariate Density Distribution

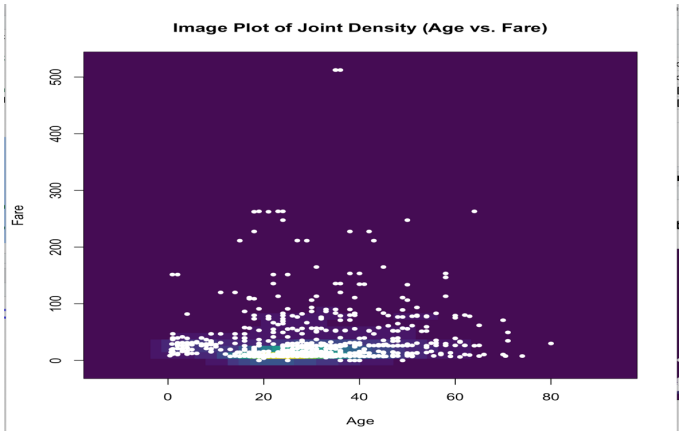
To understand the joint distribution of Age and Fare, we used the ks package in R to estimate the **bivariate kernel density** then the results are visualized

🔷 Image (Heat) Plot

```
# b. Visualize using an Image (Heat) Plot with a Viridis Color Palette
image(x = kde_result$eval.points[[1]],
      y = kde_result$eval.points[[2]],
      z = kde_result$estimate,
      col = viridis(20),
      xlab = "Age",
      ylab = "Fare",
      main = "Image Plot of Joint Density (Age vs. Fare)")

# Overlay the original data points to add context
points(kde_result$x, pch = 20, col = "white")
```

Chart:

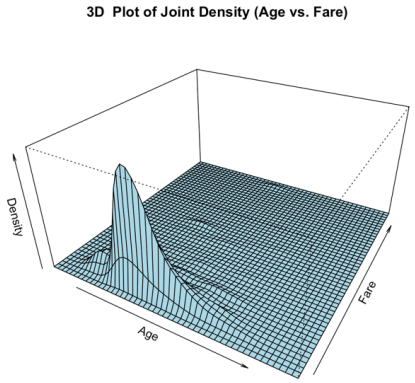


- The overall distribution is **right-skewed** on the Fare axis.
- There's **no strong linear relationship** between Age and Fare.
- Most passengers were **young to middle-aged** and traveled on **budget fares**, with a **concentrated cluster** visible in the lower-left region of the plot.

🔷 3D Perspective Plot

A three-dimensional surface plot was created to illustrate the topology of the density landscape.

```
#3d plot
persp(kde_result$eval.points[[1]],
      kde_result$eval.points[[2]],
      kde_result$estimate,
      theta = 30,         # Rotation angle (degrees)
      phi = 30,           # Elevation angle (degrees)
      expand = 0.5,       # Scaling factor for the z-axis
      col = "lightblue", # Color for the surface
      xlab = "Age",
      ylab = "Fare",
      zlab = "Density",
      main = "3D  Plot of Joint Density (Age vs. Fare)")
```

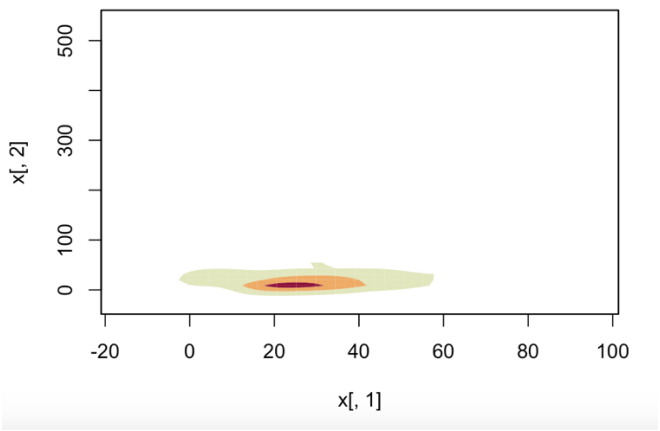


- The **majority of passengers** were **younger and paid cheaper fares**.
- There's **no strong dependency** between Age and Fare—high-density areas are localized rather than trending diagonally.
- This 3D view reinforces the findings from the heat plot: the data is **skewed toward low-fare, younger individuals**, with limited density elsewhere.

🔷 Slice and Contour Plots

```
#Slice plot
plot(kde_result, display="filled.contour2")
```

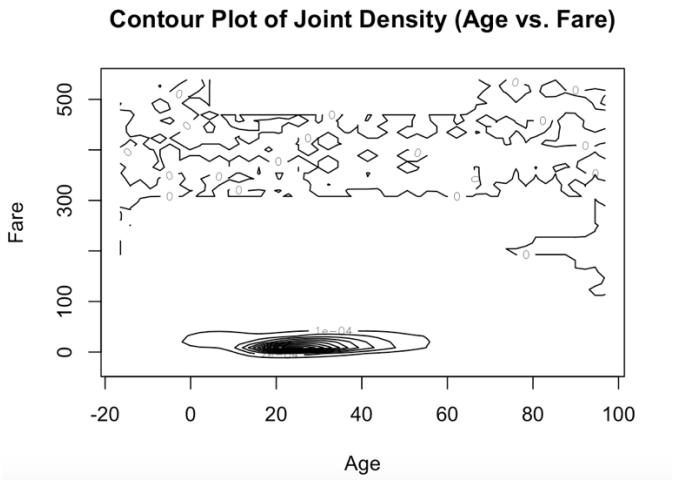
.



- The contour plot reaffirms that most passengers were **young to middle-aged adults** who paid **low to moderate fares**.
- The distribution is **unimodal** and fairly compact, with **no strong nonlinear patterns**.
- Age and Fare are **not strongly correlated**, but they do cluster in one dominant region.

🔷Contour Plot

```
#Visualize using a Contour Plot
contour(x = kde_result$eval.points[[1]],
        y = kde_result$eval.points[[2]],
        z = kde_result$estimate,
        xlab = "Age",
        ylab = "Fare",
        main = "Contour Plot of Joint Density (Age vs. Fare)")
```



- The contour plot confirms that the **majority of Titanic passengers were younger and paid lower fares**.
- The distribution is **unimodal** with a compact high-density core.
- **Age and Fare are not strongly correlated**; the relationship appears weak, with density driven more by passenger clustering in certain demographics.

3. Conclusions: Shape, Direction, and Strength

The joint distribution is heavily **right-skewed** on the Fare axis on all plot. Most passengers paid lower fares and fell within a broad age range of 20–40.

Direction: No strong **linear relationship** exists between Age and Fare. The plot shows clusters of younger passengers at low fares and some older passengers paying both low and high fares (likely due to first-class status).

Strength: The relationship between Age and Fare is **weak**. The density contours are spread out with no clear trend, suggesting little to no correlation between these two variables.

4. Bandwidth Matrix HH

The bandwidth matrix HH determines the level of smoothing in kernel density estimation. In this analysis, we computed HH using **smoothed cross-validation** via the Hscv() function from the **ks** package. This method selects an optimal bandwidth that balances bias and variance by minimizing mean integrated squared error (MISE).

The resulting HH matrix is symmetric and positive definite, ensuring a smooth and stable estimate of the joint density function. Choosing a good HH is essential:

Too small → noisy, overfitted estimate

Too large → oversmoothed, loss of structure

In this case, the automatic selection resulted in visually meaningful plots that reflected key patterns in the data.