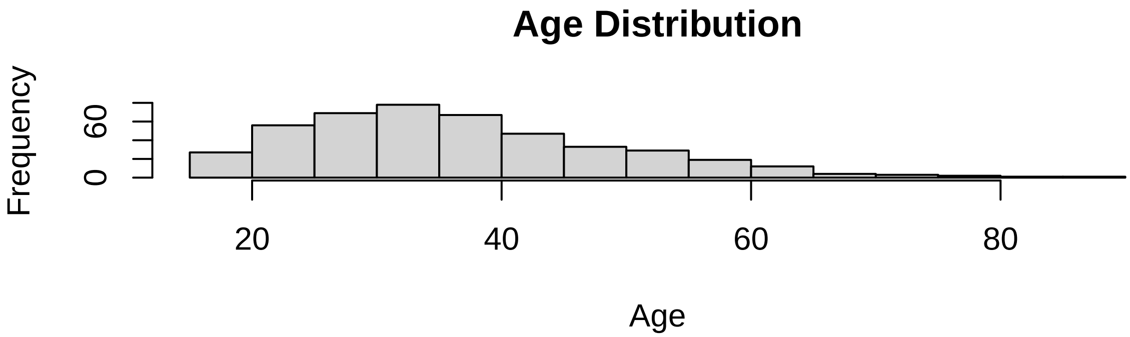
add_20241214

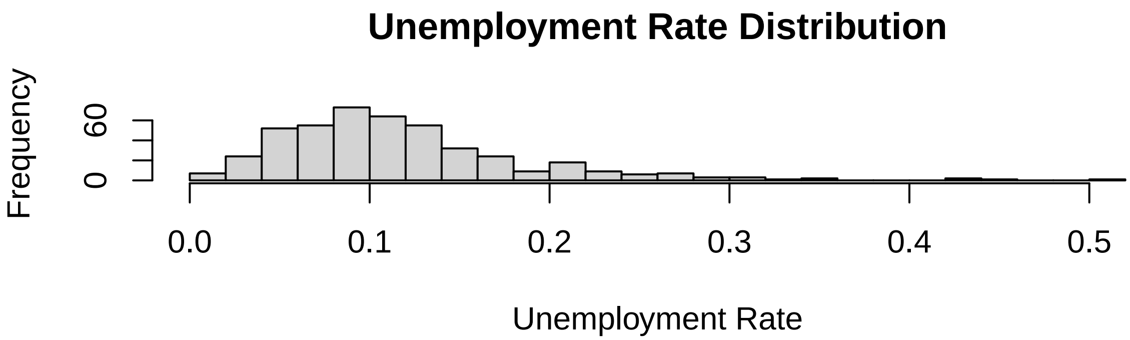
Eleanor Jiang 305002785

2024-12-14

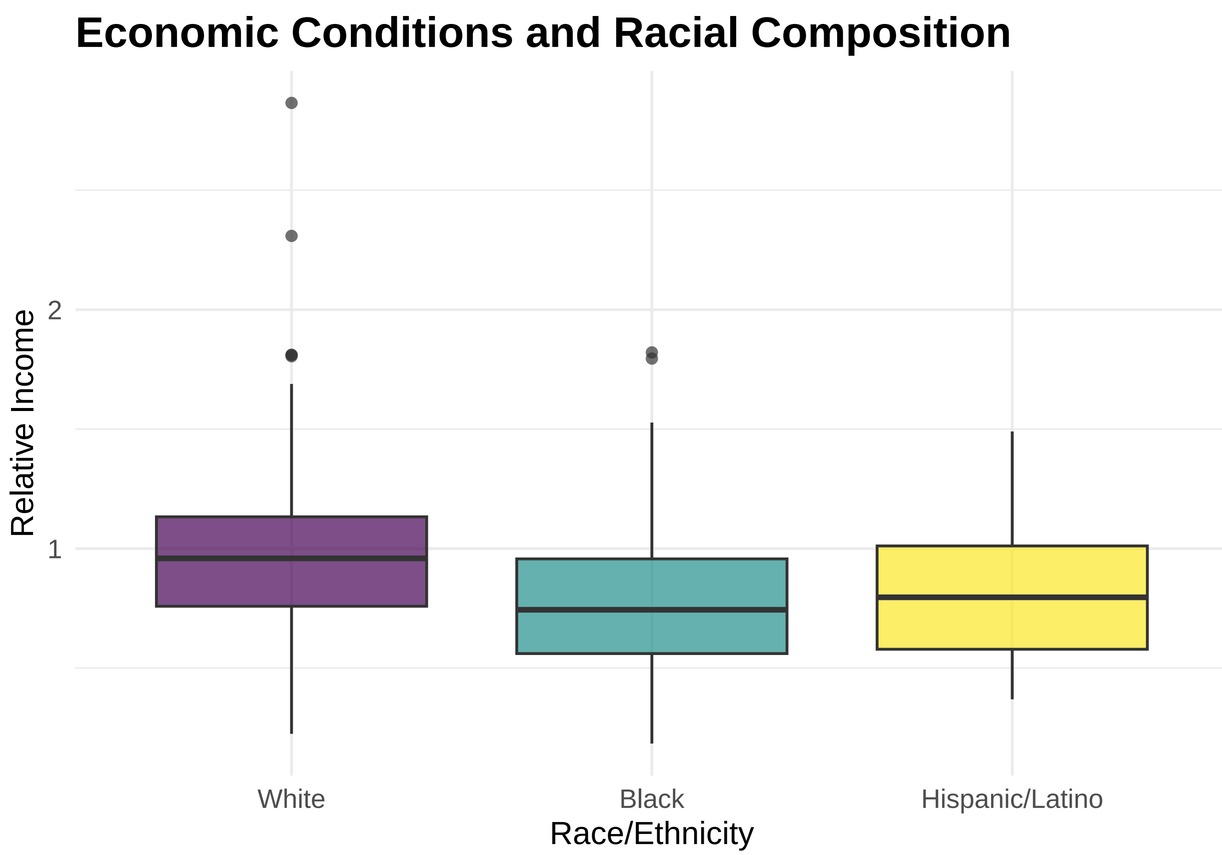
```
## Rows: 448
## Columns: 34
## $ name                 <chr> "A'donte Washington", "Aaron Rutledge", "Aaron Si~
## $ age                  <dbl> 16, 27, 26, 25, 29, 29, 22, 35, 44, 31, 76, 40, 3~
## $ gender               <fct> Male, Male, Male, Male, Male, Male, Male, Male, M~
## $ raceethnicity        <fct> Black, White, White, Hispanic/Latino, White, Whit~
## $ month                <chr> "February", "April", "March", "March", "March", "~
## $ day                  <int> 23, 2, 14, 11, 19, 7, 27, 26, 28, 7, 26, 12, 25, ~
## $ year                 <int> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2~
## $ streetaddress        <chr> "Clearview Ln", "300 block Iris Park Dr", "22nd A~
## $ city                 <chr> "Millbrook", "Pineville", "Kenosha", "South Gate"~
## $ state                <chr> "AL", "LA", "WI", "CA", "OH", "AZ", "CA", "CA", "~
## $ latitude             <dbl> 32.52958, 31.32174, 42.58356, 33.93930, 41.14857,~
## $ longitude            <dbl> -86.36283, -92.43486, -87.83571, -118.21946, -81.~
## $ state_fp             <int> 1, 22, 55, 6, 39, 4, 6, 6, 48, 26, 6, 6, 18, 18, ~
## $ county_fp            <int> 51, 79, 59, 37, 153, 13, 29, 37, 41, 81, 31, 59, ~
## $ tract_ce             <int> 30902, 11700, 1200, 535607, 530800, 111602, 700, ~
## $ geo_id               <dbl> 1051030902, 22079011700, 55059001200, 6037535607,~
## $ county_id            <int> 1051, 22079, 55059, 6037, 39153, 4013, 6029, 6037~
## $ namelsad             <chr> "Census Tract 309.02", "Census Tract 117", "Censu~
## $ lawenforcementagency <chr> "Millbrook Police Department", "Rapides Parish Sh~
## $ cause                <chr> "Gunshot", "Gunshot", "Gunshot", "Gunshot", "Guns~
## $ armed                <fct> No, No, No, Yes, No, No, Yes, Yes, Yes, Yes, Yes,~
## $ pop                  <int> 3779, 2769, 4079, 4343, 6809, 4682, 5027, 5238, 4~
## $ share_white          <dbl> 60.5, 53.8, 73.8, 1.2, 92.5, 7.0, 50.8, 8.6, 14.6~
## $ share_black          <dbl> 30.5, 36.2, 7.7, 0.6, 1.4, 7.7, 0.3, 0.2, 17.7, 7~
## $ share_hispanic       <dbl> 5.6, 0.5, 16.8, 98.8, 1.7, 79.0, 44.2, 84.1, 66.3~
## $ p_income             <dbl> 28375, 14678, 25286, 17194, 33954, 15523, 25949, ~
## $ h_income             <int> 51367, 27972, 45365, 48295, 68785, 20833, 58068, ~
## $ county_income        <int> 54766, 40930, 54930, 55909, 49669, 53596, 48552, ~
## $ comp_income          <dbl> 0.9379359, 0.6834107, 0.8258693, 0.8638144, 1.384~
## $ county_bucket        <int> 3, 2, 2, 3, 5, 1, 4, 4, 2, 3, 4, 5, 4, 3, 1, 3, 1~
## $ nat_bucket           <int> 3, 1, 3, 3, 4, 1, 4, 4, 1, 2, 3, 5, 2, 2, 1, 3, 2~
## $ pov                  <dbl> 14.1, 28.8, 14.6, 11.7, 1.9, 58.0, 17.2, 12.2, 37~
## $ urate                <dbl> 0.09768638, 0.06572379, 0.16629314, 0.12482727, 0~
## $ college              <dbl> 0.16850951, 0.11140236, 0.14731227, 0.05013293, 0~
```

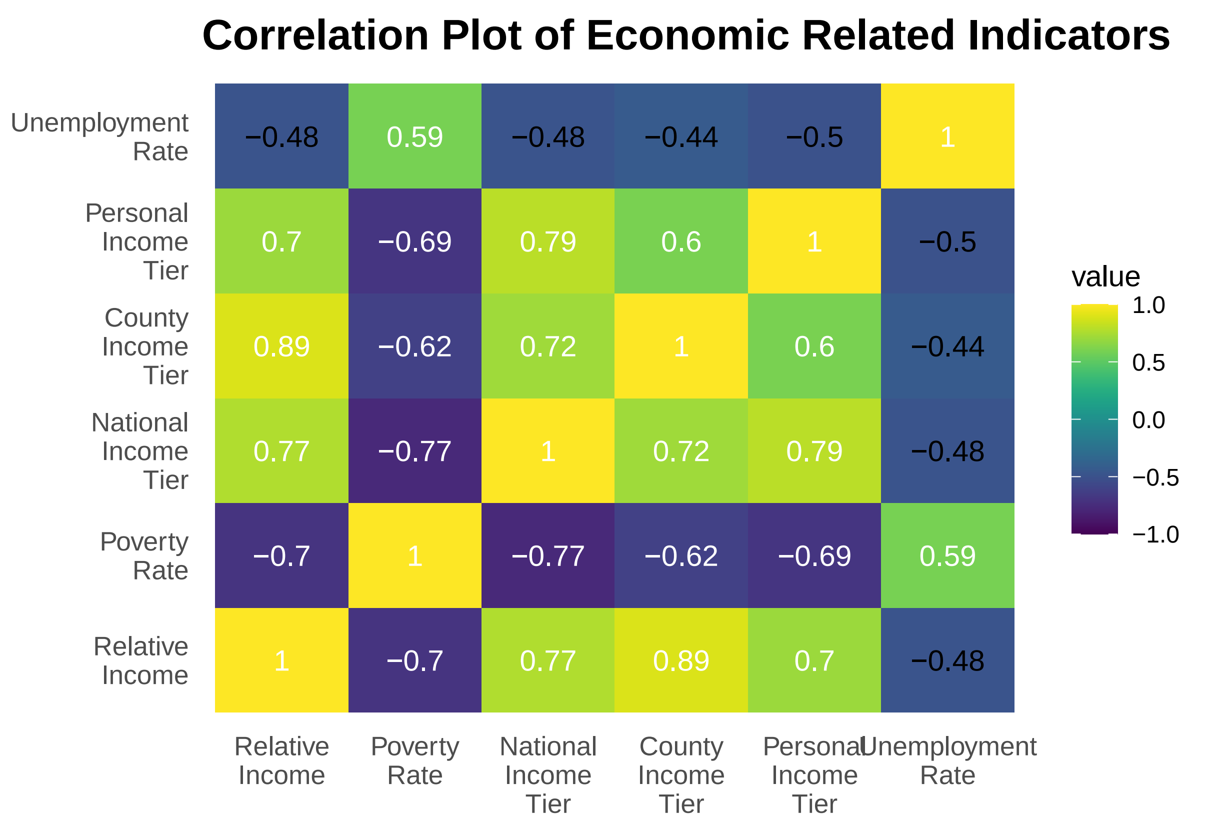
EDA





The demographic analysis shows that the deceased had a mean age in their 30s, with the age range extending to over 80. The unemployment rate in their residential tracts averaged 10%, reaching a maximum of 51%. The gender distribution was heavily skewed, with 428 males and 20 females.





Due to sample size limitations, the racial composition analysis was restricted to three main groups: White

(52%), Black, and Hispanic/Latino, with Asian and Native American groups (each <5%) with less than 30 people being excluded from the analysis. For our **research question 1**, it is clear that a multinomial logistic regression is proficient in this case. The box plot reveals distinct income disparities across racial groups, suggesting relative income as a potentially significant predictor. Analysis of economic indicators showed high correlations among most variables, with tract-level unemployment rate being the exception. Based on this correlation analysis, we selected unemployment rate and one additional economic indicator for the final model.

```r
library(nnet)
library(pander)

# Data preprocessing
rq1_data$raceethnicity <- droplevels(rq1_data$raceethnicity[
  !rq1_data$raceethnicity %in% c("Asian/Pacific Islander", "Native American")
])

# Model fitting
rq1_model <- multinom(
  raceethnicity ~ comp_income + urate + college + pop + age + share_black*pov,
  data = rq1_data
)
```

```
## # weights:  30 (18 variable)
## initial  value 476.797733
## iter  10 value 345.193804
## iter  20 value 325.104151
## final  value 325.086555
## converged
```

```r
# Model summary and statistics
summary_model <- summary(rq1_model)
coef <- summary_model$coefficients
se <- summary_model$standard.errors
z_scores <- coef / se
p_values <- 2 * (1 - pnorm(abs(z_scores)))

summary_table <- data.frame(
  Estimate = as.vector(coef),
  `Std.Error` = as.vector(se),
  `z value` = as.vector(z_scores),
  `Pr(>|z|)` = as.vector(p_values)
)

# Display using pander
library(pander)
pander(summary_table)
```

Estimate	Std.Error	z.value	Pr...z..
0.8195	0.0001909	4293	0
1.267	0.0003199	3963	0
-1.1	0.0001554	-7080	0
-0.6648	0.0002588	-2569	0
2.255	3.446e-05	65457	0
1.327	4.421e-05	30014	0
2.501	3.931e-05	63608	0

Estimate	Std.Error	z.value	Pr...z..
-1.182	9.798e-05	-12068	0
-7.902e-05	5.11e-05	-1.546	0.122
-8.661e-06	5.639e-05	-0.1536	0.8779
-0.0426	0.008538	-4.989	6.061e-07
-0.07276	0.01134	-6.416	1.399e-10
0.04169	0.01369	3.045	0.002328
0.01525	0.02649	0.5756	0.5649
-0.01369	0.01341	-1.021	0.3074
0.04703	0.01441	3.263	0.001101
0.0002493	0.0004544	0.5488	0.5832
-0.001571	0.001032	-1.523	0.1277

```r
# Model accuracy
predicted_classes <- predict(rq1_model, rq1_data)
accuracy <- mean(predicted_classes == rq1_data$raceethnicity)
cat("Model Accuracy:", round(accuracy, 4), "\n")
```

```
## Model Accuracy: 0.7028
```

```r
# AIC comparison
null_model <- multinom(raceethnicity ~ 1, data = rq1_data)
```

```
## # weights:  6 (2 variable)
## initial  value 476.797733
## final  value 425.765085
## converged
```

```r
cat("Null Model AIC:", null_model$AIC, "\n")
```

```
## Null Model AIC: 855.5302
```

```r
cat("Full Model AIC:", rq1_model$AIC, "\n")
```

```
## Full Model AIC: 686.1731
```

The difference in AIC between the null model and rq1_model is:855.5302 - 686.1731 = 169.3571. This substantial reduction in AIC (169.36 points) indicates that rq1_model provides a much better fit compared to the null model, as lower AIC values suggest better model fit. A difference greater than 10 points is typically considered strong evidence for the superior model. The model also has an accuracy of 0.7, indicating a fair performance of inference of our outcome variable.