New York City Crime Report Analysis

FIRE STARS - Princi Sharma & Maitri Bhatt

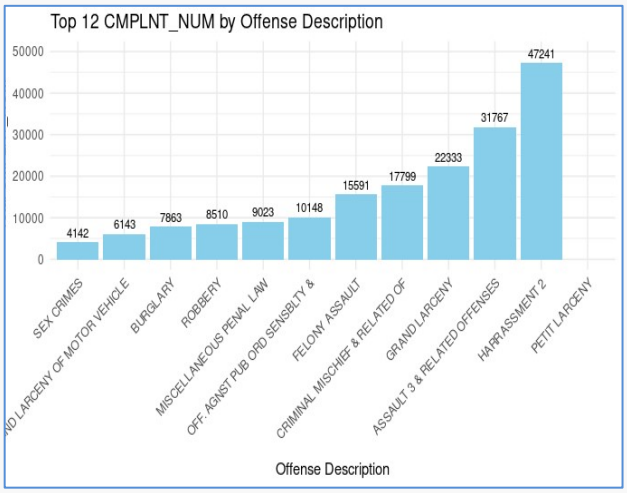
```
data <-
read.csv("NYPD_Complaint_Data_Current__Year_To_Date_.csv",stringsAsFactors =
FALSE)


#Column Descriptions
```

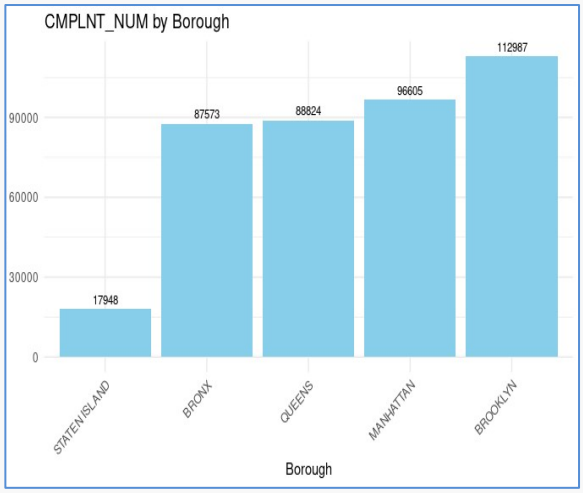
Crime_Column_Description

Column	Description
CMPLNT_NUM	Randomly generated persistent ID for each complaint
CMPLNT_FR_DT	Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)
CMPLNT_FR_TM	Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists)
CMPLNT_TO_DT	Ending date of occurrence for the reported event, if exact time of occurrence is unknown
CMPLNT_TO_TM	Ending time of occurrence for the reported event, if exact time of occurrence is unknown
RPT_DT	Date event was reported to police
KY_CD	Three digit offense classification code
OFNS_DESC	Description of offense corresponding with key code
PD_CD	Three digit internal classification code (more granular than Key Code)
PD_DESC	Description of internal classification corresponding with PD code (more granular than Offense Description)
CRM_ATPT_CPTD_CD	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
JURIS_DESC	Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc.
BORO_NM	The name of the borough in which the incident occurred
ADDR_PCT_CD	The precinct in which the incident occurred
LOC_OF_OCCUR_DESC	Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of
PREM_TYP_DESC	Specific description of premises; grocery store, residence, street, etc.
PARKS_NM	Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)
HADEVELOPT	Name of NYCHA housing development of occurrence, if applicable
X_COORD_CD	X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Y_COORD_CD	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Latitude	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

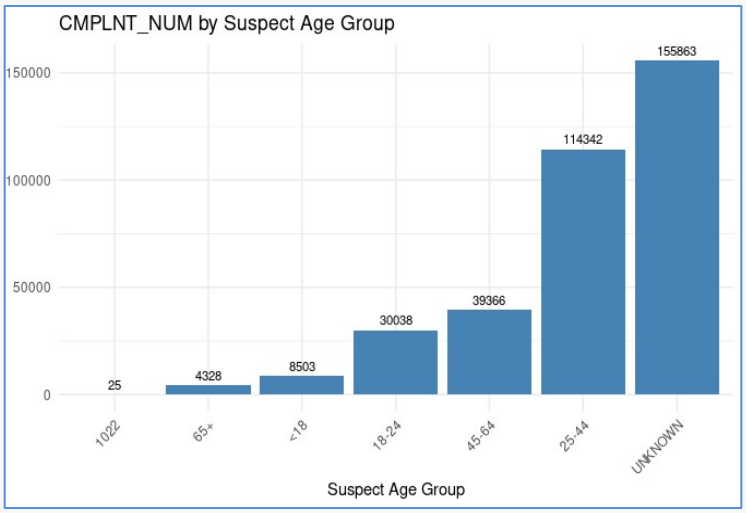
Visualizations: Exploratory Data Analysis

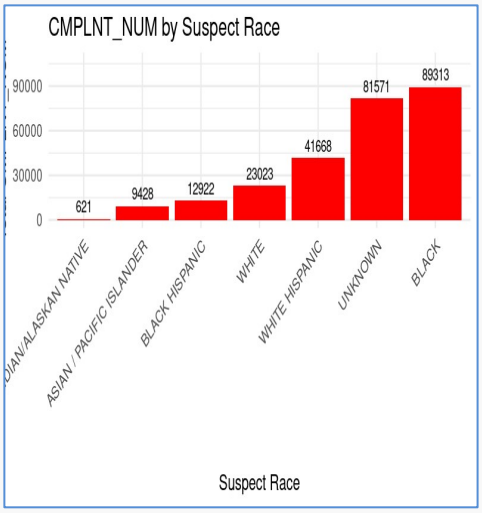
```
library(ggplot2)
```



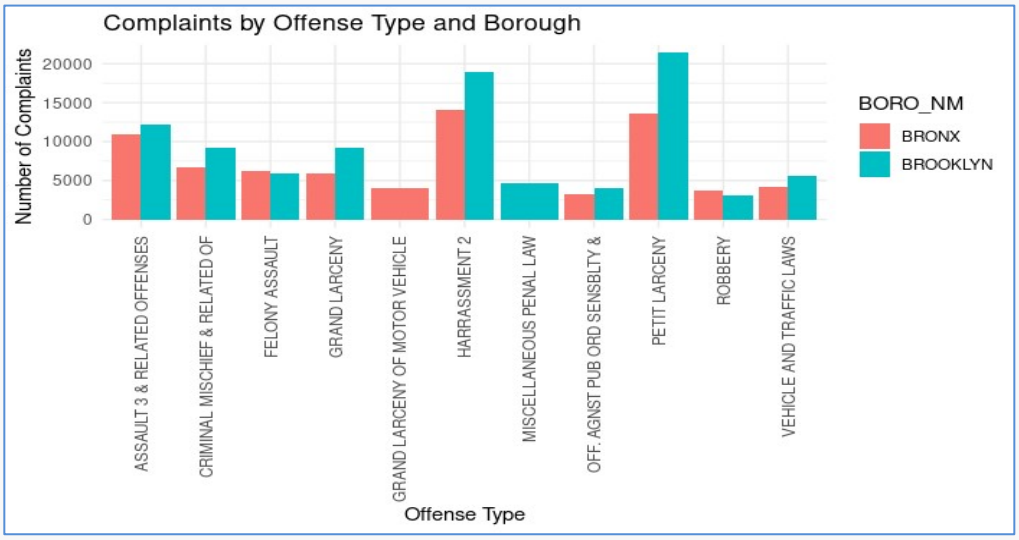


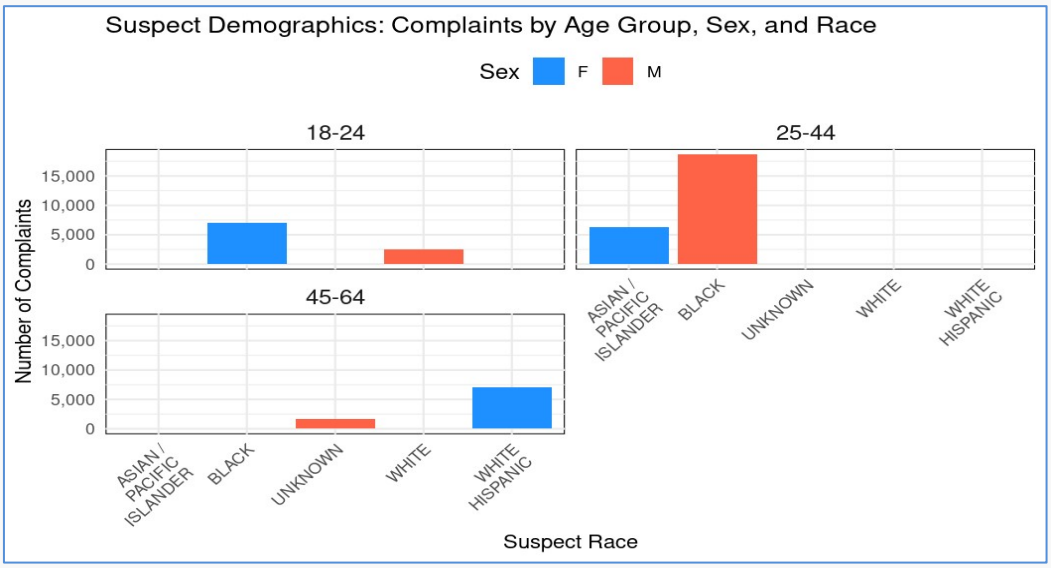
```
# Group data by OFNS_DESC and calculate count
```

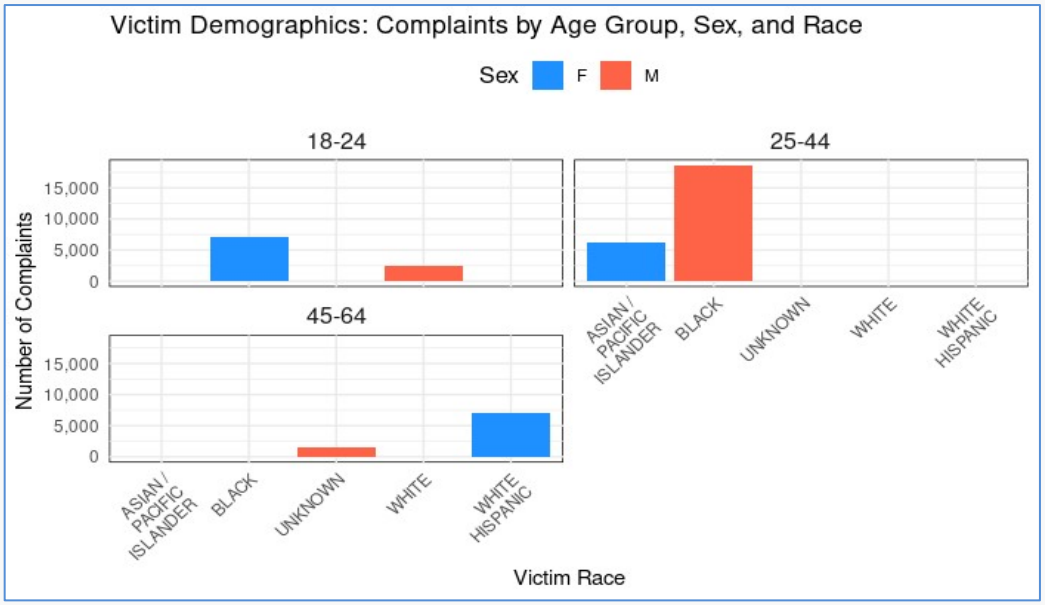




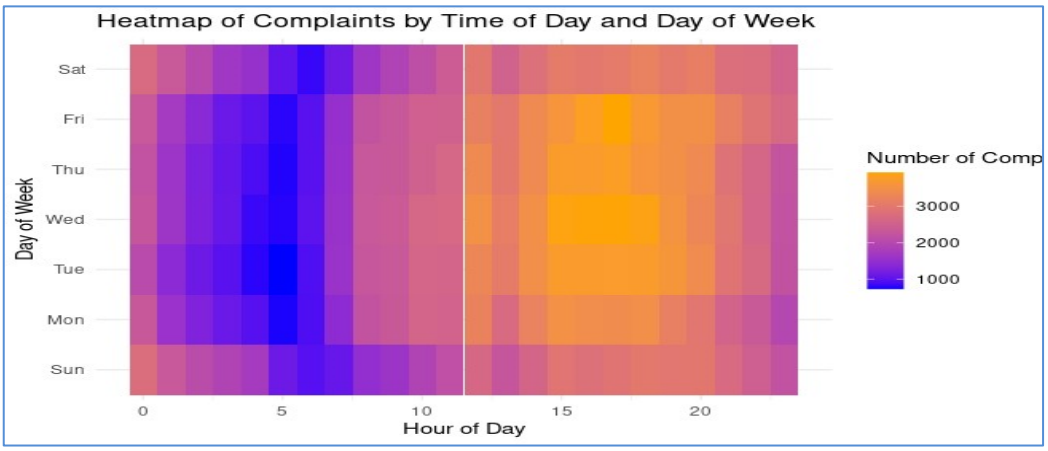

```
# Select the top 5 crime types
```

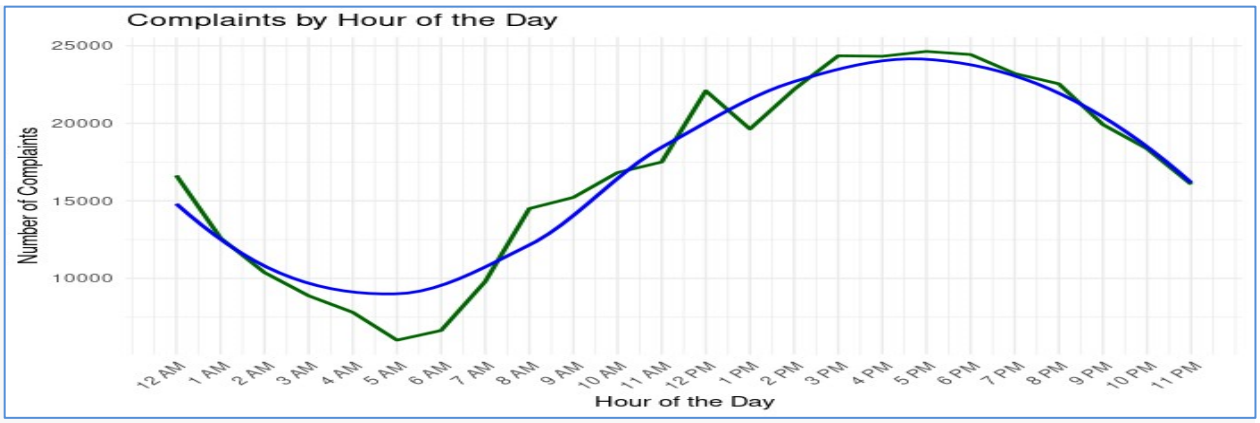


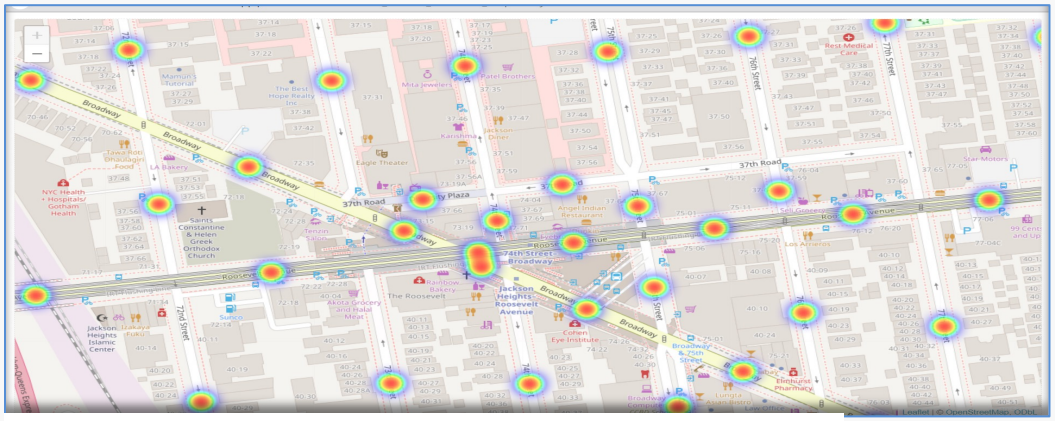





```
# Plots for the top 5 crimes in BROOKLYN
```

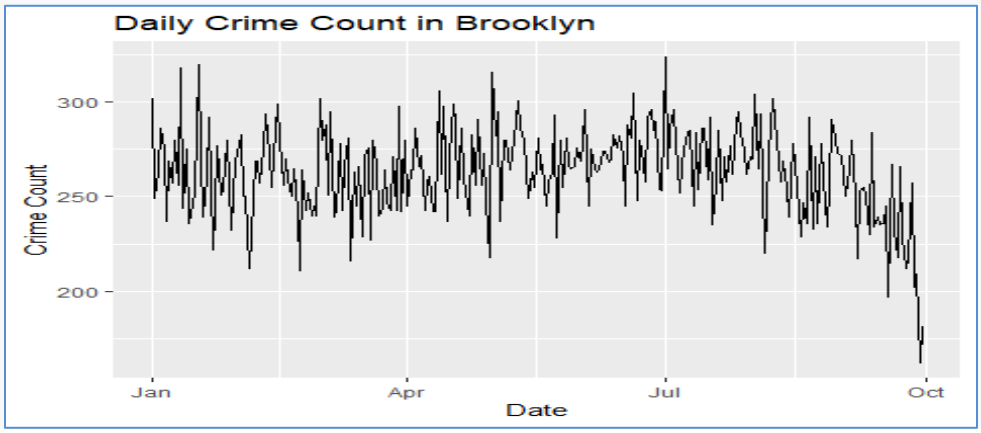






Season <chr>	Total_Crimes <int>	Avg_Temperature <dbl>
Fall	7106	65.68407
Spring	24383	48.08055
Summer	24846	69.07555
Winter	15390	36.64256

Time Series Analysis



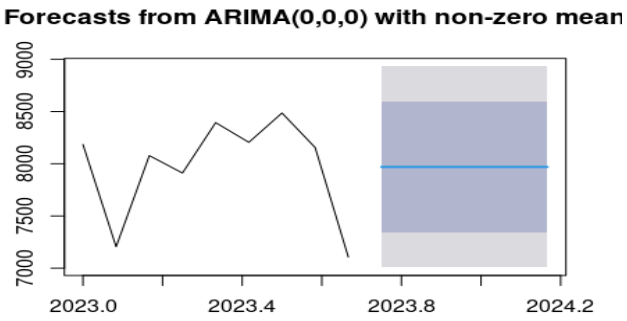
The time series graph shows daily crime counts in Brooklyn with notable variability and an apparent downward trend or decrease in crime occurrences towards the end of the year.

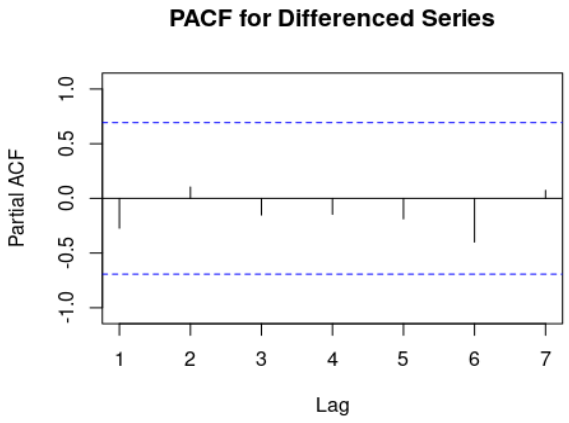
```r
# Fit the ARIMA model
print(forecast)
```

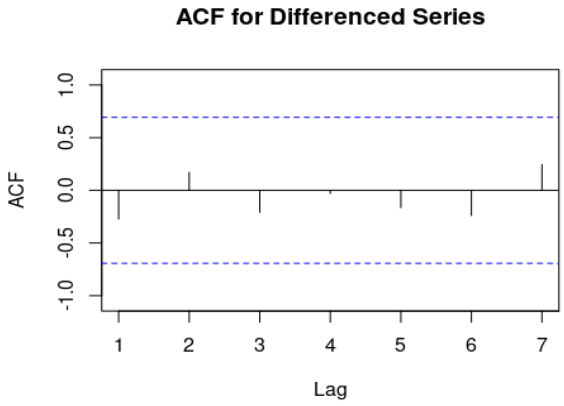
```
##    Point       Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## Oct 2023       7969.444 7340.215 8598.674 7007.122 8931.767
## Nov 2023       7969.444 7340.215 8598.674 7007.122 8931.767
## Dec 2023       7969.444 7340.215 8598.674 7007.122 8931.767
## Jan 2024       7969.444 7340.215 8598.674 7007.122 8931.767
## Feb 2024       7969.444 7340.215 8598.674 7007.122 8931.767
## Mar 2024       7969.444 7340.215 8598.674 7007.122 8931.767
```

The forecast output indicates a constant point forecast value of 7969.444 for each month, with the 80% and 95% prediction intervals suggesting increasing uncertainty over time, yet remaining symmetrical around the forecasted mean.

```r
# ACF and PACF plots for residuals
```







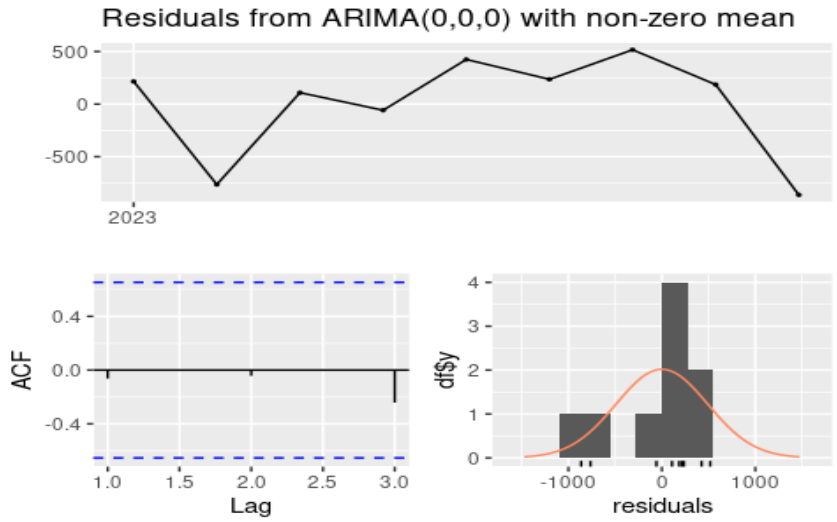
The forecast from the ARIMA(0,0,0) model with a non-zero mean suggests a static forecast projecting the historical mean into the future, with increasing forecast uncertainty indicated by the widening confidence intervals.

The Augmented Dickey-Fuller test results indicate extremely high p-values (0.99) for both the original and differenced crime time series, strongly suggesting that the series is non-stationary even after differencing.

The PACF plot for the differenced series shows no significant partial autocorrelations at any lag, which suggests that the differencing has sufficiently removed any autoregressive components and the series may not require further AR terms.

```
fit <- auto.arima(crime_ts, seasonal = FALSE)

## Series: crime_ts
## ARIMA(0,0,0) with non-zero mean
##
## Coefficients:
##             mean
##       7969.4444
## s.e.   154.3043
##
## sigma^2 = 241071:  log likelihood = -68.01
## AIC=140.02   AICc=142.02   BIC=140.41
##
## Training set error measures:
##                       ME     RMSE      MAE         MPE     MAPE MASE
## Training set 1.010673e-13 462.9099 374.7407 -0.3589992 4.863137  NaN
##                      ACF1
## Training set -0.06305476
```

The ARIMA(0,0,0) model with a non-zero mean suggests a simplistic approach with a mean forecast value of 7969.44, but the relatively large standard error of the mean and high sigma^2 value indicate considerable uncertainty and variability in the model's predictions.



While the residuals do not exhibit autocorrelation, the distribution of residuals suggests the ARIMA(0,0,0) model may not fully capture the underlying process of the data, and there might be room for improvement by considering additional parameters or transformations.

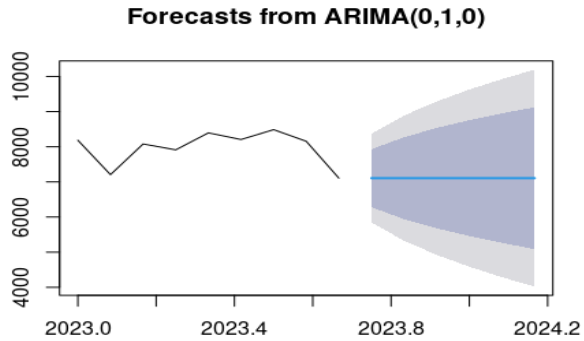
```
## Ljung-Box test
## data:  Residuals from ARIMA(0,0,0) with non-zero mean
## Q* = 1.0401, df = 3, p-value = 0.7916
## Model df: 0.   Total lags used: 3
```

The Ljung-Box test on the residuals from the ARIMA(0,0,0) model with a non-zero mean suggests no significant autocorrelation at the first 3 lags, as indicated by a high p-value of 0.7916.

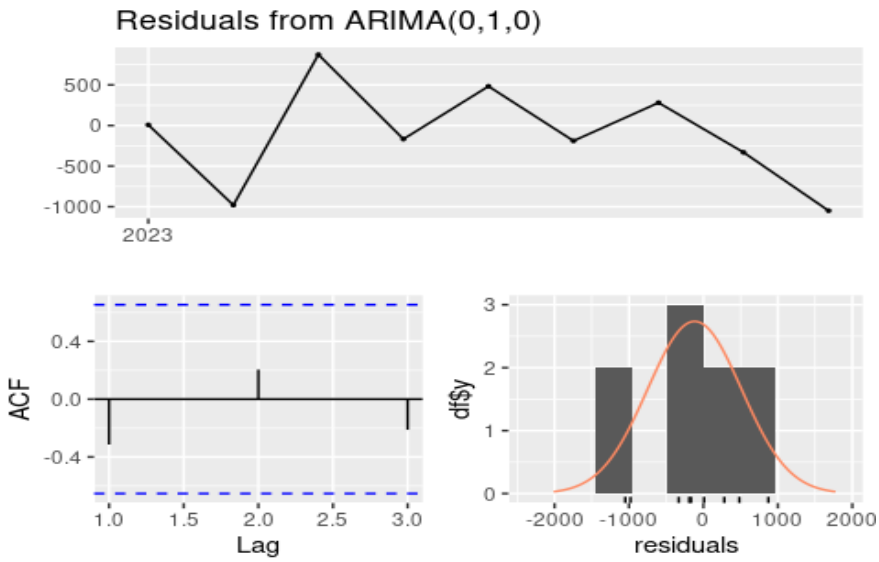
```
# Fit an ARIMA(0,1,0) model
fit2 <- Arima(crime_ts, order=c(0,1,0))

# Summary of the model
summary(fit2)
```

```
## Series: crime_ts
## ARIMA(0,1,0)
## sigma^2 = 413552:  log likelihood = -63.08
## AIC=128.16   AICc=128.83   BIC=128.24
## Training set error measures:
##                     ME     RMSE      MAE       MPE     MAPE MASE       ACF1
## Training set -118.9794 606.3019 484.5761 -1.873842 6.312193  NaN -
0.3137135
```

The ARIMA(0,1,0) model applied to the crime time series data yields a relatively high error, with significant mean absolute error (MAE) and mean absolute percentage error (MAPE), indicating the model's predictions deviate notably from the actual values.



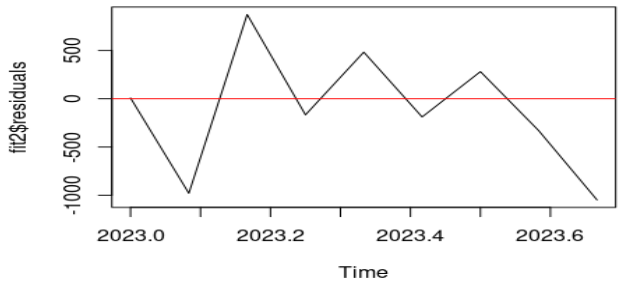
The forecast plot from the ARIMA(0,1,0) model shows a flat forecast line with increasing uncertainty over time, as depicted by the widening confidence intervals, indicating constant future values based on a random walk model.



The composite plot indicates that the residuals from the ARIMA(0,1,0) model display no significant autocorrelation and are approximately normally distributed, which suggests the model fits the data well.

```
## Ljung-Box test
## data:  Residuals from ARIMA(0,1,0)
## Q* = 2.5557, df = 3, p-value = 0.4653
## Model df: 0.   Total lags used: 3
```

The Ljung-Box test results indicate that with a p-value of 0.4653 for residuals from the ARIMA(0,1,0) model, there is no significant evidence of autocorrelation at lag 3, suggesting the residuals are essentially random (white noise).



The plot shows the residuals from the fit2 ARIMA model, which are fluctuating significantly around the zero line, suggesting the model may not be adequately capturing all patterns in the data.

```
# Compare AIC, AICc, and BIC
AIC(fit)

## [1] 140.0165
```

```r
AICc(fit)
```

```
## [1] 142.0165
```

```r
BIC(fit)
```

```
## [1] 140.4109
```

```r
AIC(fit2)
```

```
## [1] 128.1632
```

```r
AICc(fit2)
```

```
## [1] 128.8298
```

```r
BIC(fit2)
```

```
## [1] 128.2426
```

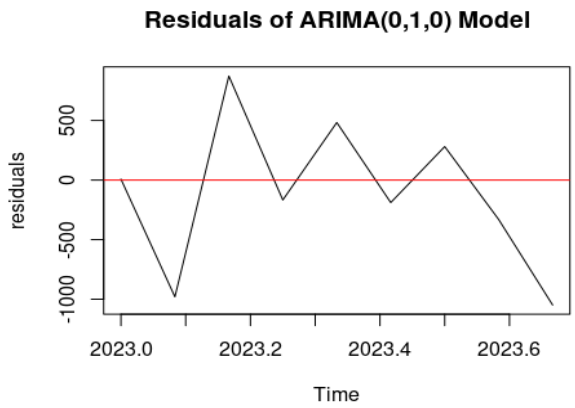
```r
# Compare error measures (RMSE, MAE, etc.)
accuracy(fit)
```

```
##                       ME     RMSE      MAE         MPE     MAPE MASE
## Training set 1.010673e-13 462.9099 374.7407 -0.3589992 4.863137  NaN
##                      ACF1
## Training set -0.06305476
```

```r
accuracy(fit2)
```

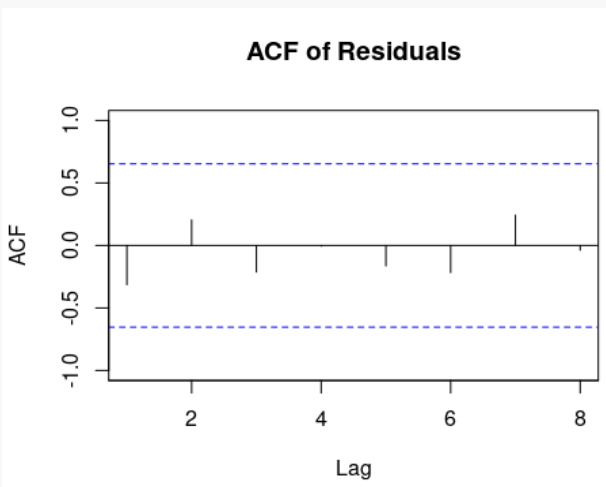
```
##                      ME     RMSE      MAE       MPE     MAPE MASE       ACF1
## Training set -118.9794 606.3019 484.5761 -1.873842 6.312193  NaN -0.3137135
```

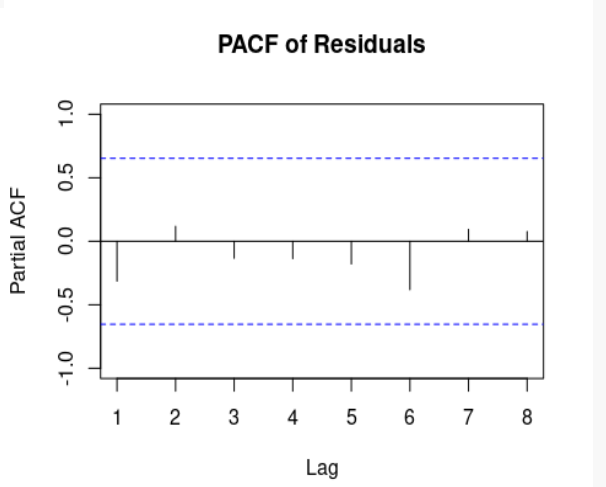
While fit2 appears to be the better model based on information criteria (AIC, AICc, BIC), the accuracy measures suggest that fit provides forecasts closer to the actual values with less error and bias. This discrepancy implies that while fit2 may be a more parsimonious model, fit could be more reliable for forecasting. The choice between the two models should consider the trade-off between complexity (favoring fit2) and forecasting accuracy (favoring fit).s



The diagram displays the residuals of an ARIMA(0,1,0) model over time, showing fluctuations around the zero line without a discernible pattern, which is a characteristic of a potentially well-fitting model.

```
# ACF and PACF plots for residuals
```





The diagram is an Autocorrelation Function (ACF) plot of residuals, indicating that there is no significant autocorrelation at lags 1 through 8 as all autocorrelations are within the confidence interval bounds, suggesting that the residuals may be random (white noise).

The diagram is a Partial Autocorrelation Function (PACF) plot of residuals, showing that there are no significant partial autocorrelations at lags 1 through 8, as all bars are within the confidence interval bounds, suggesting that the residuals could be white noise.

```
##  Box-Ljung test
## data:  residuals
## X-squared = NA, df = 20, p-value = NA

# Linear regression between Temperature and Total Crimes
crime_temp_model <- lm(Total_Crimes ~ Temperature, data = crime_by_temp)
```

```
Call:
lm(formula = Total_Crimes ~ Temperature, data = crime_by_temp)

Residuals:
     Min       1Q   Median       3Q      Max
-1437.47  -542.74   -75.91   360.65  2105.22

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  518.870    301.407   1.721   0.0907 .
Temperature   14.069      5.576   2.523   0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 758.2 on 56 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1021,    Adjusted R-squared:  0.08603
F-statistic: 6.365 on 1 and 56 DF,  p-value: 0.0145
```

The linear regression analysis indicates a statistically significant positive relationship between temperature and total crimes, with each unit increase in temperature associated with an increase of approximately 14.069 in total crimes, as evidenced by a p-value of 0.0145.

The change in Total_Crimes for a one-unit change in Temperature. Only 10% Temperature affects the crime rates.

```
# Random Forest model
model <- randomForest(SUSP_SEX ~ ., data = data_selected, ntree = 50)
```

```
##
## Call:
##  randomForest(formula = SUSP_SEX ~ ., data = data_selected, ntree = 50)
##                Type of random forest: classification
##                      Number of trees: 50
## No. of variables tried at each split: 1
##
##         OOB estimate of  error rate: 27.1%
## Confusion matrix:
##   F     M     U class.error
## F 0 11902  1419   1.0000000
## M 0 36443  4191   0.1031402
## U 0  1927 15843   0.1084412
```

The random forest classifier with 50 trees has an overall out-of-bag error rate of 27.1%, and it completely misclassified the 'F' category, which suggests significant model bias and poor predictive performance, particularly for the 'F' class.

```
# Confusion Matrix
confusionMatrix <- table(data_selected$SUSP_SEX, predictions)
```

```
##    predictions
##         F     M     U
##   F     0 11901  1420
##   M     0 36440  4194
##   U     0  1913 15857
```

The classification results show a complete misclassification of the 'F' category, with all 'F' predicted as 'M' or 'U', and a relatively lower misclassification rate for 'M' and 'U', indicating the model is significantly biased and ineffective at correctly predicting the 'F' category.

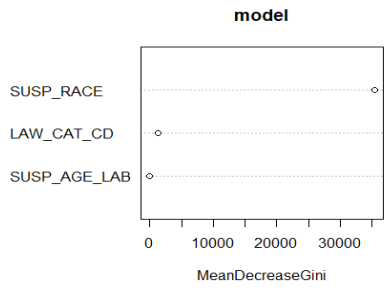
```
# Tune the model
tuned_model <- randomForest(SUSP_SEX ~ ., data = data_selected, ntree = 100,
mtry = 3)
```

```
model <- randomForest(SUSP_SEX ~ ., data = data_selected, ntree = 100)
## Call:
##  randomForest(formula = SUSP_SEX ~ ., data = data_selected, ntree = 100)
##                Type of random forest: classification
```

```
##                       Number of trees: 100
## No. of variables tried at each split: 1
##
##         OOB estimate of  error rate: 26.29%
## Confusion matrix:
##   F      M     U class.error
## F 0  39778  4919   1.0000000
## M 0 129938 15924   0.1091717
## U 0   7348 60639   0.1080795
```

The random forest model with 100 trees has an out-of-bag (OOB) error rate of 26.29% and completely misclassifies all 'F' observations while showing more than 10% class error for both 'M' and 'U' categories, indicating a model with considerable prediction bias and room for improvement.

```
# Feature importance
importance <- importance(model)
varImpPlot(model)
```



The model's variable importance plot indicates that 'SUSP_RACE' is the most influential predictor in determining the outcome, with 'LAW_CAT_CD' and 'SUSP_AGE_LAB' having a lesser, yet substantial, impact on the model's decisions.

```
#ANOVA MODELS

#Fit ANOVA model for Crime & Temperature
anova_model <- aov(Total_Crimes ~ Season , data = seasonal_crime_trends)
```

```
##             Df    Sum Sq  Mean Sq
## Season       3 213082835 71027612
```

The ANOVA model indicates that 'Season' explains a significant portion of the variance in 'Total_Crimes' with a sum of squares of 213,082,835 spread across three degrees of freedom, averaging a mean square of 71,027,612.

```
# Fit ANOVA model
anova_model <- aov(CMPLNT_NUM ~ VIC_AGE_GROUP_REDUCE, data = filtered_data)
##   term                    df   sumsq  meansq statistic p.value
```

```
##   <chr>                <dbl>   <dbl>   <dbl>     <dbl>   <dbl>
## 1 VIC_AGE_GROUP_REDUCE     2 7.88e11 3.94e11    0.0250   0.975
## 2 Residuals            38378 6.05e17 1.58e13   NA       NA
```

The ANOVA model for 'CMPLNT_NUM' and 'VIC_AGE_GROUP_REDUCE' suggests that there is no significant difference in means between the age group categories, as indicated by a p-value of 0.975, and the majority of variance is explained by the residuals.

```
# Post hoc analysis (Tukey HSD test) for pairwise comparisons
posthoc <- TukeyHSD(anova_model)
print(posthoc)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = CMPLNT_NUM ~ VIC_AGE_GROUP_REDUCE, data =
filtered_data)
##
## $VIC_AGE_GROUP_REDUCE
##                    diff       lwr      upr     p adj
## 25-44-18-24  -203.0589 -130026.0 129619.9 0.9999926
## 65+-18-24   14989.4037 -176210.3 206189.1 0.9815607
## 65+-25-44   15192.4627 -145511.2 175896.1 0.9733000
```

The Tukey post-hoc test reveals no statistically significant differences in the number of complaints among the age groups 18-24, 25-44, and 65+, as indicated by the very high p-values and large confidence intervals that encompass zero.

```
# Fit ANOVA model
anova_model <- aov(CMPLNT_NUM ~ LAW_CAT_CD  + SUSP_RACE + SUSP_SEX +
Temperature, data = brooklyn_data)
print(anova_model)
```

```
## Call:
##    aov(formula = CMPLNT_NUM ~ LAW_CAT_CD + SUSP_RACE + SUSP_SEX +
##     Temperature, data = brooklyn_data)
##
## Terms:
##                    LAW_CAT_CD    SUSP_RACE     SUSP_SEX  Temperature
## Sum of Squares  2.080406e+13 1.606742e+15 2.055368e+14 8.218029e+17
## Deg. of Freedom             2            6            2            1
##                     Residuals
## Sum of Squares  3.203426e+17
## Deg. of Freedom        71713
##
## Residual standard error: 2113530
## Estimated effects may be unbalanced
```

```
##   VIC_AGE_GROUP_REDUCE Total_Crimes
##   <chr>                       <int>
```

```
## 1 18-24                          6280
## 2 25-44                         28296
## 3 65+                            3805
## 4 <NA>                          33344
```

The ANOVA results indicate that there is no significant effect of the victim's age group on the sum of squares of the complaint numbers, with a very high p-value (0.975) suggesting no statistical difference across the groups when compared to the variability within the residuals.