Analysis of Arrest Data in South Bureau of LAPD

Presented By –

- **Shashank Sangu**

ABSTRACT

This project report provides a detailed analysis of arrest data in Los Angeles, focusing on insights to shape law enforcement strategies and policies. Using data from the Los Angeles Police Department (LAPD), the report examines various aspects such as the relationship between age and gender of arrestees, the correlation between race and types of charges, the likelihood of facing felony charges based on demographic factors, and temporal patterns in arrests. Key findings reveal that female arrestees tend to be younger than males, suggesting a need for tailored intervention programs. Additionally, there is a notable correlation between race and charges, indicating opportunities for targeted law enforcement efforts. Demographic factors like age, gender, and location significantly influence the likelihood of facing felony charges, emphasizing the need for fairer practices and resource allocation. While daily arrest counts show no clear trends, a predictive model helps forecast future figures for better resource distribution. The report suggests implementing support programs based on gender, strategic law enforcement measures informed by racial and demographic factors, and continuous improvement of predictive models for enhanced crime prevention. Overall, this analysis provides valuable insights for stakeholders to collaborate on creating safer and fairer communities in Los Angeles.

Introduction

In the vibrant city of Los Angeles, the complex interplay of law enforcement and the criminal justice system influences communities and individuals daily. Delving into the details of arrest data is crucial for making informed decisions and crafting policies that prioritize public safety and fairness. This report conducts a thorough examination of arrest data in the Los Angeles region, addressing critical research questions to reveal insights that can shape law enforcement approaches and policy interventions.

Why?

We have opted this dataset for gaining insights into arrest incidents in Los Angeles. It helps us understand crime patterns, law enforcement activities, and trends in public safety. With the LAPD switching to a new Records Management System to meet the FBI's NIBRS mandate, having comprehensive historical data is crucial for developing and testing the new system. Despite possible inaccuracies from manually transcribing paper reports, this dataset is a fundamental resource for analyzing crime data and ensuring the successful rollout of the new reporting system. Ultimately, it supports law enforcement efforts and community safety initiatives.

About the Dataset:

The LAPD dataset contains detailed records of arrest incidents in Los Angeles from May 24, 2020, to April 10, 2024. With 278,000 rows and 25 columns, each row represents a unique arrest event, offering insight into law enforcement activities. Managed by LAPD Open Data, the dataset falls under Public Safety, crucial for analyzing crime trends in the city.

Key attributes of the dataset include the date of arrest, location specified through latitude and longitude coordinates, and various demographic and incident-specific details such as the age, gender, race of the arrestee, the type of offense, and the outcome of the arrest. Despite potential inaccuracies due to data transcription from original arrest reports, the dataset offers valuable insights into crime patterns, trends, and law enforcement responses. With a committed update frequency of weekly refreshes, the dataset remains current and relevant, supporting ongoing efforts to promote public safety and ensure transparency in policing practices within Los Angeles.

The variables in the dataset are classified as shown below:

Categorical variables:

- Report ID
- Report Type
- Area ID
- Area Name
- Reporting District
- Sex Code
- Descent Code
- Charge Group Code
- Charge Group Description
- Arrest Type Code
- Charge
- Charge Description
- Disposition Description
- Address
- Cross Street
- Booking Location
- Booking Location Code

Numerical variables:

- Age
- LAT (Latitude)
- LON (Longitude)

Date Time variables:

- Arrest Date
- Time
- Booking Date

Presentation of Data and Data Analysis-

Data Cleaning and Transformation:

Before proceeding with the analysis, the raw arrest data underwent several cleaning steps to ensure accuracy and consistency:

Adding New Column:

New columns were introduced to the dataset, namely Felony, Race, and Age Category, to streamline the analysis process and organize data into more manageable categories.

Felony Column:

To simplify the analysis, the initial dataset included a wide array of charge types. To streamline this complexity, the charges were grouped into two categories: Felony and Non-Felony. Charges classified as felonies were designated as 'Felony', while all other charges were categorized as ‘non-Felony'.

Data Type Conversion:

Data types underwent standardization to ensure consistency and compatibility throughout the dataset. This process included converting variables like dates or categorical data into appropriate formats suitable for analysis.

Data Analysis -

After completing the data cleaning phase, the dataset was prepared for detailed analysis. The subsequent key analyses were conducted.

Key Variables Used for Analysis –

Here's a brief overview of the key variables used in the analysis:

1. **Age:** Age describes the arrestee's age at the time of the arrest. It helps uncover age-related patterns in criminal activity and sheds light on the demographics of those who are arrested.

2. **Sex Code:** The arrestee's gender is indicated by their sex code, which usually classifies them as male or female. Policies and intervention programs that target gender groups can be better informed by an understanding of the gender differences in arrest rates.

3. **Area Name:** The area name indicates the precise location of the arrest. To better allocate resources and develop policing strategies, it assists in identifying patterns and trends in arrests throughout various Los Angeles communities or regions.

4. **Felony:** Felony is a binary variable that indicates whether the arrestee's charge is considered a felony or not. This variable is essential to comprehending the seriousness of the charges and determining how demographic factors affect the probability of felonies.

5. **Charge Group: T**he arrestee's charges are grouped into more comprehensive categories based on their type. Charge-related patterns and trends can be more easily interpreted thanks to this variable, which streamlines the study by combining similar charges together.

6. **Report Type:** Report type describes the kind of report that was created in relation to the arrest, including incident, citation, and arrest reports. It gives background information on the nature of the arrest and the ensuing court cases.

7. **Age Category:** Arrestees are categorized by age using pre-established categories, such as Young, Middle-aged, and Old. By classifying people based on age demographics, this variable makes analysis easier to do and makes it possible to compare and find age-related trends in arrests.

Together, these important factors offer insightful information about the demographics, geographic distribution, seriousness of charges, and age-related patterns of arrests in Los Angeles City's South Bureau.

Research questions for Analysis -

a. Is there a significant difference in the average age between male and female arrestees?
b. Is there any significant relationship between Race and type of charge they are being arrested for?
c. We want to know the likelihood of an arrestee being "Felony" based on the demographical factors like Sex Code, Area Name, Race and Age.
d. We want to analyze whether there is any temporal pattern of number of arrests and build a model to predict daily number of arrests in Los Angeles.

1. Is there a significant difference in the average age between male and female arrestees?

To begin with, we would like to know if we find a significant difference, whether gender may be a significant factor associated with differences in the average age of arrestees. This information could be valuable for understanding demographic patterns in arrests and potentially informing policies or interventions targeting specific age groups within male and female populations. Since there are only two categories in *Sex Code* variable, we will use independent samples t-test (left-tailed).

Setting up the hypothesis:

Null Hypothesis (H0): The average age of female arrestees is greater than or equal to the average age of male arrestees.

H0: $\mu_{\text{female}} \geq \mu_{\text{male}}$

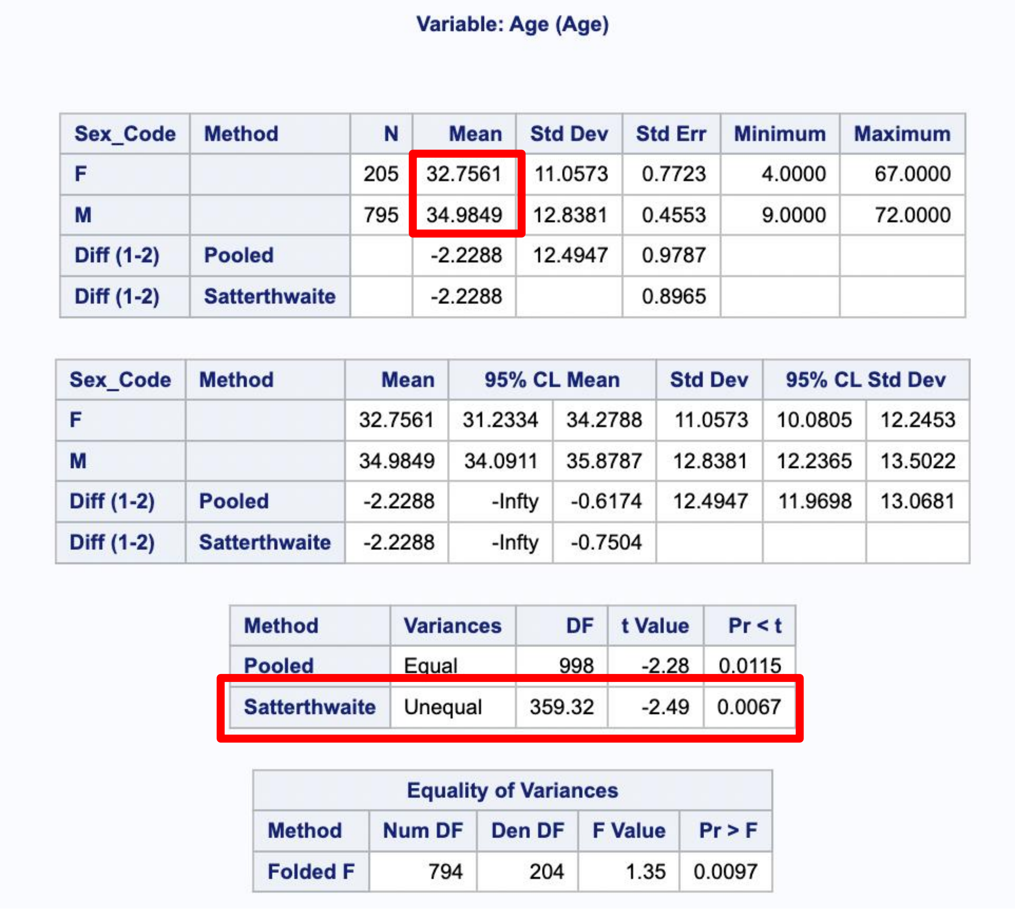
where μ_{male} is the population mean age of male arrestees and μ_{female} is the population mean age of female arrestees.

Alternative Hypothesis (H1): The average age of female arrestees is less than the average age of male arrestees.

H1: $\mu_{\text{female}} < \mu_{\text{male}}$

Not assuming equal variances and we will use one tailed t-test:

The test results are as follow:



Variable: Age (Age)

Sex_Code	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
F		205	32.7561	11.0573	0.7723	4.0000	67.0000
M		795	34.9849	12.8381	0.4553	9.0000	72.0000
Diff (1-2)	Pooled		-2.2288	12.4947	0.9787		
Diff (1-2)	Satterthwaite		-2.2288		0.8965		

Sex_Code	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
F		32.7561	31.2334	34.2788	11.0573	10.0805	12.2453
M		34.9849	34.0911	35.8787	12.8381	12.2365	13.5022
Diff (1-2)	Pooled	-2.2288	-Infty	-0.6174	12.4947	11.9698	13.0681
Diff (1-2)	Satterthwaite	-2.2288	-Infty	-0.7504			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	998	-2.28	0.0115
Satterthwaite	Unequal	359.32	-2.49	0.0067

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	794	204	1.35	0.0097

It is evident that the *p-value (0.0067)* is less than α=0.05, so we reject the null hypothesis and conclude that female arrestees tend to be younger, on average, compared to male arrestees.

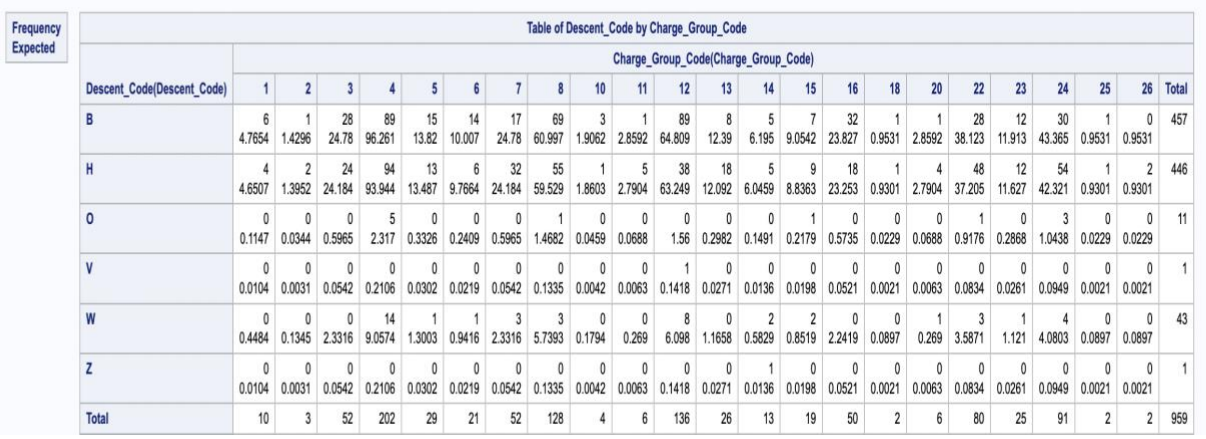
Female arrestees tend to be younger, on average, compared to male arrestees.

2. **Is there any significant relationship between Race and type of charge they are being arrested for?**

Here, we want to know whether there is a significant relationship between the race of arrestee and the type of charge they have been arrested for:

***H*₀: Variables are independent.**
***H*₁: Variables are dependent.**

We performed Chi-Square test to check the dependency between two categorical variables and obtained the following results:

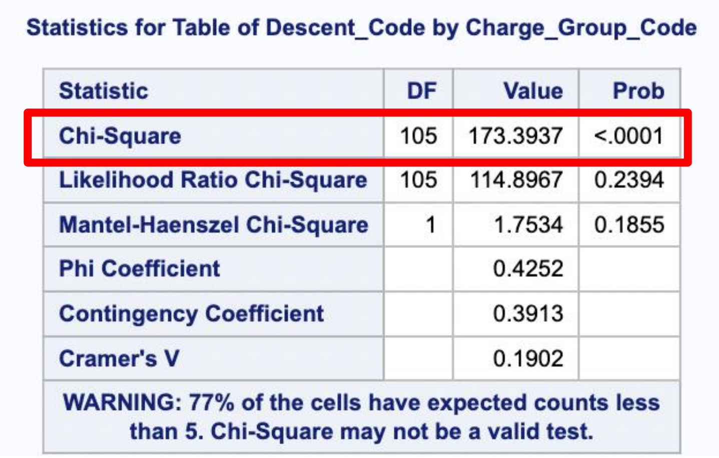


Frequency
Expected

Table of Descent_Code by Charge_Group_Code

Charge_Group_Code(Charge_Group_Code)

Descent_Code(Descent_Code)	1	2	3	4	5	6	7	8	10	11	12	13	14	15	16	18	20	22	23	24	25	26	Total
B	6 4.7654	1 1.4296	28 24.78	89 96.261	15 13.82	14 10.007	17 24.78	69 60.997	3 1.9062	1 2.8592	89 64.809	8 12.39	5 6.195	7 9.0542	32 23.827	1 0.9531	1 2.8592	28 38.123	12 11.913	30 43.365	1 0.9531	0 0.9531	457
H	4 4.6507	2 1.3952	24 24.184	94 93.944	13 13.487	6 9.7664	32 24.184	55 59.529	1 1.8603	5 2.7904	38 63.249	18 12.092	5 6.0459	9 8.8363	18 23.253	1 0.9301	4 2.7904	48 37.205	12 11.627	54 42.321	1 0.9301	2 0.9301	446
O	0 0.1147	0 0.0344	0 0.5965	5 2.317	0 0.3326	0 0.2409	0 0.5965	1 1.4682	0 0.0459	0 0.0688	0 1.56	0 0.2982	0 0.1491	1 0.2179	0 0.5735	0 0.0229	0 0.0688	1 0.9176	0 0.2868	3 1.0438	0 0.0229	0 0.0229	11
V	0 0.0104	0 0.0031	0 0.0542	0 0.2106	0 0.0302	0 0.0219	0 0.0542	0 0.1335	0 0.0042	0 0.0063	1 0.1418	0 0.0271	0 0.0136	0 0.0198	0 0.0521	0 0.0021	0 0.0063	0 0.0834	0 0.0261	0 0.0949	0 0.0021	0 0.0021	1
W	0 0.4484	0 0.1345	0 2.3316	14 9.0574	1 1.3003	1 0.9416	3 2.3316	3 5.7393	0 0.1794	0 0.269	8 6.098	0 1.1658	2 0.5829	2 0.8519	0 2.2419	0 0.0897	1 0.269	3 3.5871	1 1.121	4 4.0803	0 0.0897	0 0.0897	43
Z	0 0.0104	0 0.0031	0 0.0542	0 0.2106	0 0.0302	0 0.0219	0 0.0542	0 0.1335	0 0.0042	0 0.0063	0 0.1418	0 0.0271	1 0.0136	0 0.0198	0 0.0521	0 0.0021	0 0.0063	0 0.0834	0 0.0261	0 0.0949	0 0.0021	0 0.0021	1
Total	10	3	52	202	29	21	52	128	4	6	136	26	13	19	50	2	6	80	25	91	2	2	959



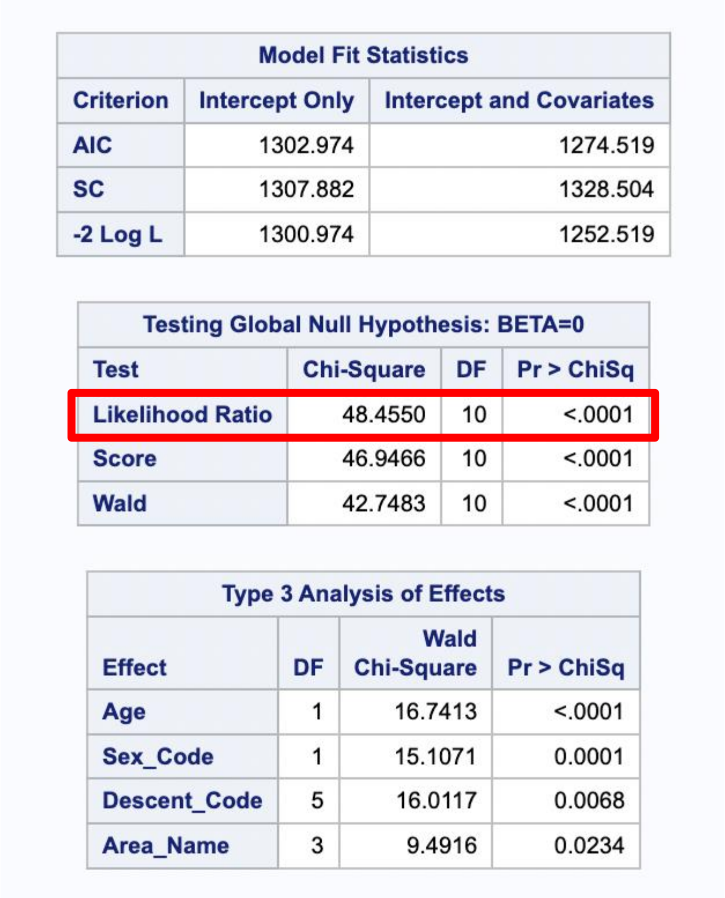
Statistics for Table of Descent_Code by Charge_Group_Code

Statistic	DF	Value	Prob
Chi-Square	105	173.3937	<.0001
Likelihood Ratio Chi-Square	105	114.8967	0.2394
Mantel-Haenszel Chi-Square	1	1.7534	0.1855
Phi Coefficient		0.4252	
Contingency Coefficient		0.3913	
Cramer's V		0.1902	
WARNING: 77% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

From the above results, it is evident that *p-value* for Chi-Square statistic is less that 0.05. hence, we can conclude that charge type is dependent on the race of an arrestee.

3. **We want to know the likelihood of an arrestee being "Felony" based on the demographical factors like Sex Code, Area Name, Race and Age.**

For this analysis, we used logistic regression.
The following are the test results:



Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1302.974	1274.519
SC	1307.882	1328.504
-2 Log L	1300.974	1252.519

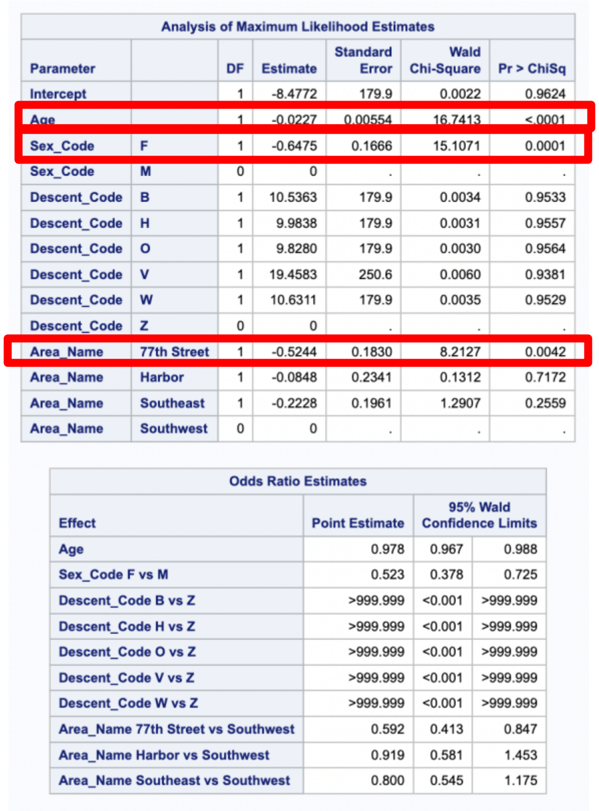
Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	48.4550	10	<.0001
Score	46.9466	10	<.0001
Wald	42.7483	10	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	16.7413	<.0001
Sex_Code	1	15.1071	0.0001
Descent_Code	5	16.0117	0.0068
Area_Name	3	9.4916	0.0234

For the overall model, we check the likelihood ratio Chi-sq test. However, *p-value* is very small so we can reject the null hypothesis and conclude the overall model to be significant.



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-8.4772	179.9	0.0022	0.9624
Age		1	-0.0227	0.00554	16.7413	<.0001
Sex_Code	F	1	-0.6475	0.1666	15.1071	0.0001
Sex_Code	M	0	0	.	.	.
Descent_Code	B	1	10.5363	179.9	0.0034	0.9533
Descent_Code	H	1	9.9838	179.9	0.0031	0.9557
Descent_Code	O	1	9.8280	179.9	0.0030	0.9564
Descent_Code	V	1	19.4583	250.6	0.0060	0.9381
Descent_Code	W	1	10.6311	179.9	0.0035	0.9529
Descent_Code	Z	0	0	.	.	.
Area_Name	77th Street	1	-0.5244	0.1830	8.2127	0.0042
Area_Name	Harbor	1	-0.0848	0.2341	0.1312	0.7172
Area_Name	Southeast	1	-0.2228	0.1961	1.2907	0.2559
Area_Name	Southwest	0	0	.	.	.

Odds Ratio Estimates

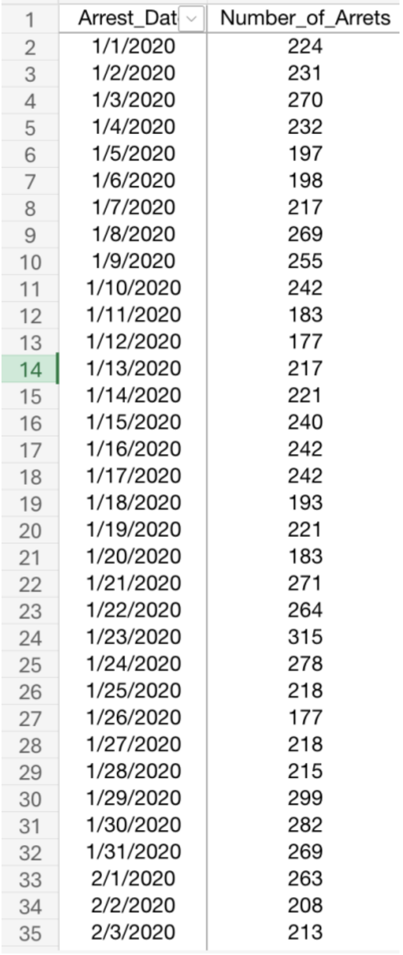
Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.978	0.967	0.988
Sex_Code F vs M	0.523	0.378	0.725
Descent_Code B vs Z	>999.999	<0.001	>999.999
Descent_Code H vs Z	>999.999	<0.001	>999.999
Descent_Code O vs Z	>999.999	<0.001	>999.999
Descent_Code V vs Z	>999.999	<0.001	>999.999
Descent_Code W vs Z	>999.999	<0.001	>999.999
Area_Name 77th Street vs Southwest	0.592	0.413	0.847
Area_Name Harbor vs Southwest	0.919	0.581	1.453
Area_Name Southeast vs Southwest	0.800	0.545	1.175

For individual variables, we check the Wald Chi-sq test. *p-values* for the variables *Age, Female and Area Name 77th Street* are less than 0.05. We can conclude that these three variables are having significant effect of likelihood of an arrestee being "Felony".

Given the age and area, the odds of a female being arrested as a felony is 0.523 times less likely than a male. Also, given the gender and age, the odds of a person from 77th Street being arrested as a felony is 0.592 times less likely than a person from Southwest.

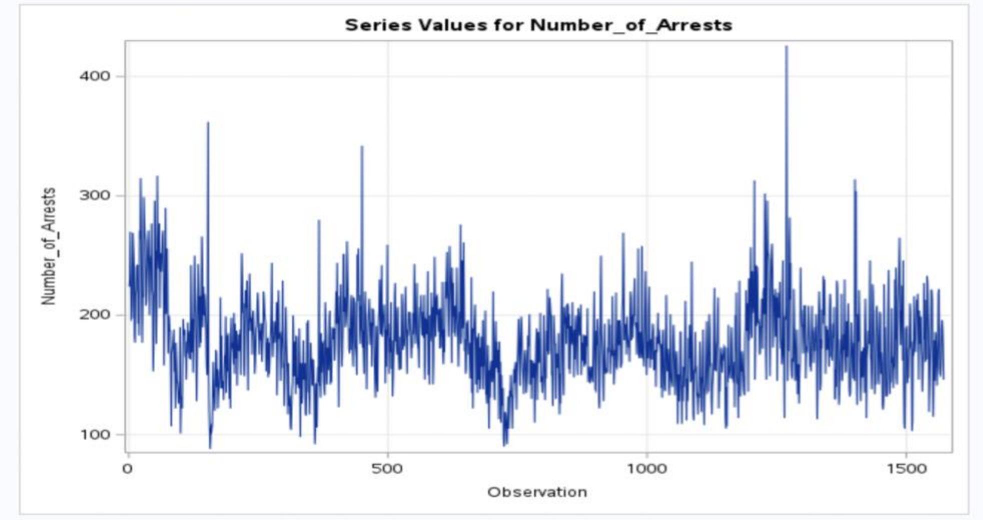
4. **We want to analyze whether there is any temporal pattern of number of arrests from the past 4 years of data which is from 2020 till the date and build a model to predict daily number of arrests in Los Angeles.**

   The daily number of arrests over the last 4 years for the entire Los Angeles city have been calculated based on arrest date and below is the sample data:

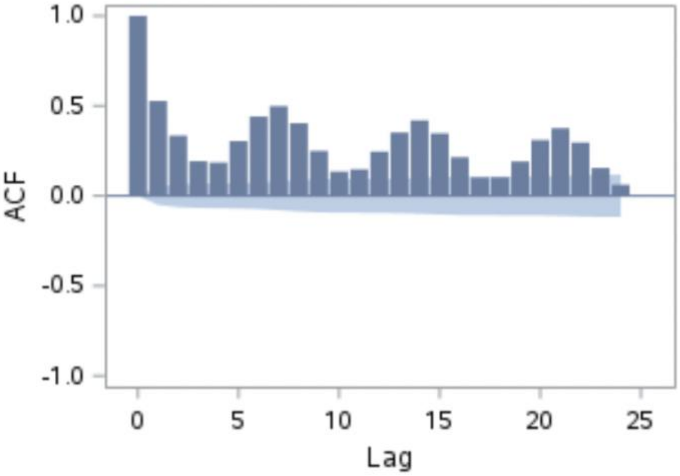


1	Arrest_Dat	Number_of_Arrets
2	1/1/2020	224
3	1/2/2020	231
4	1/3/2020	270
5	1/4/2020	232
6	1/5/2020	197
7	1/6/2020	198
8	1/7/2020	217
9	1/8/2020	269
10	1/9/2020	255
11	1/10/2020	242
12	1/11/2020	183
13	1/12/2020	177
14	1/13/2020	217
15	1/14/2020	221
16	1/15/2020	240
17	1/16/2020	242
18	1/17/2020	242
19	1/18/2020	193
20	1/19/2020	221
21	1/20/2020	183
22	1/21/2020	271
23	1/22/2020	264
24	1/23/2020	315
25	1/24/2020	278
26	1/25/2020	218
27	1/26/2020	177
28	1/27/2020	218
29	1/28/2020	215
30	1/29/2020	299
31	1/30/2020	282
32	1/31/2020	269
33	2/1/2020	263
34	2/2/2020	208
35	2/3/2020	213

Firstly, we will do time series analysis to check whether there is a pattern for the daily number of arrests and perform all randomness checks.

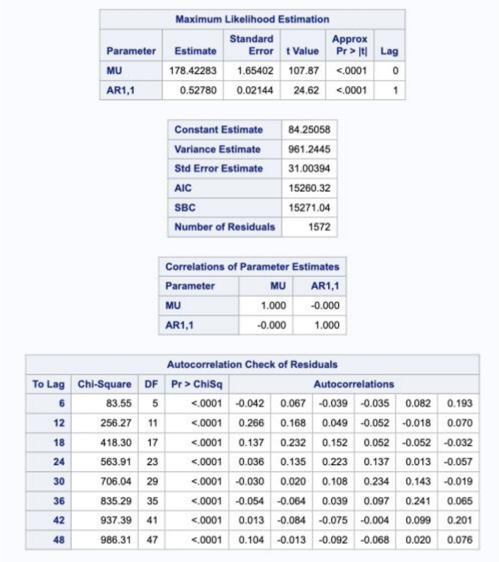


As we can see, there is no specific pattern in the above time-series plot. Thus, the underlying daily number of arrests is random.



From the above ACF plot, there are many autocorrelation bars falling out of 95% limits, which confirms non-randomness.

Now, we will try to fit a simple ARIMA model.



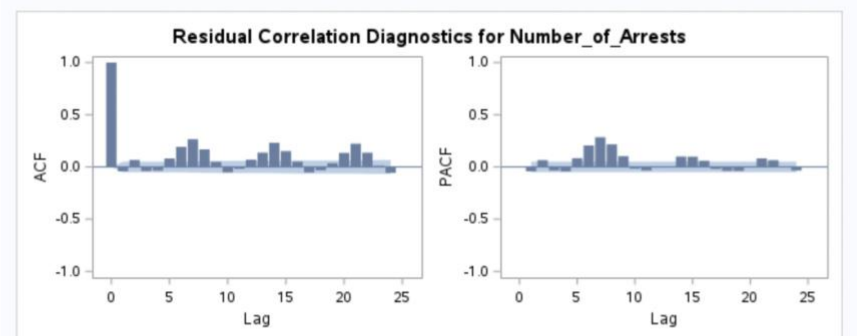
Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	178.42283	1.65402	107.87	<.0001	0
AR1,1	0.52780	0.02144	24.62	<.0001	1

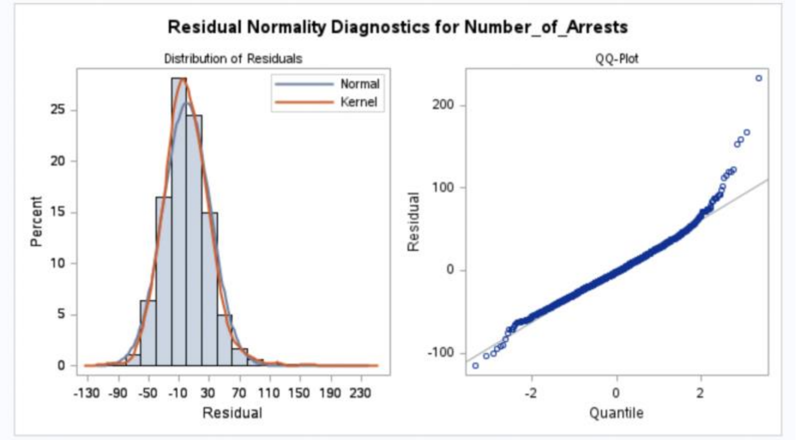
Constant Estimate	84.25058
Variance Estimate	961.2445
Std Error Estimate	31.00394
AIC	15260.32
SBC	15271.04
Number of Residuals	1572

Correlations of Parameter Estimates		
Parameter	MU	AR1,1
MU	1.000	-0.000
AR1,1	-0.000	1.000

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	83.55	5	<.0001	-0.042	0.067	-0.039	-0.035	0.082	0.193
12	256.27	11	<.0001	0.266	0.168	0.049	-0.052	-0.018	0.070
18	418.30	17	<.0001	0.137	0.232	0.152	0.052	-0.052	-0.032
24	563.91	23	<.0001	0.036	0.135	0.223	0.137	0.013	-0.057
30	706.04	29	<.0001	-0.030	0.020	0.108	0.234	0.143	-0.019
36	835.29	35	<.0001	-0.054	-0.064	0.039	0.097	0.241	0.065
42	937.39	41	<.0001	0.013	-0.084	-0.075	-0.004	0.099	0.201
48	986.31	47	<.0001	0.104	-0.013	-0.092	-0.068	0.020	0.076

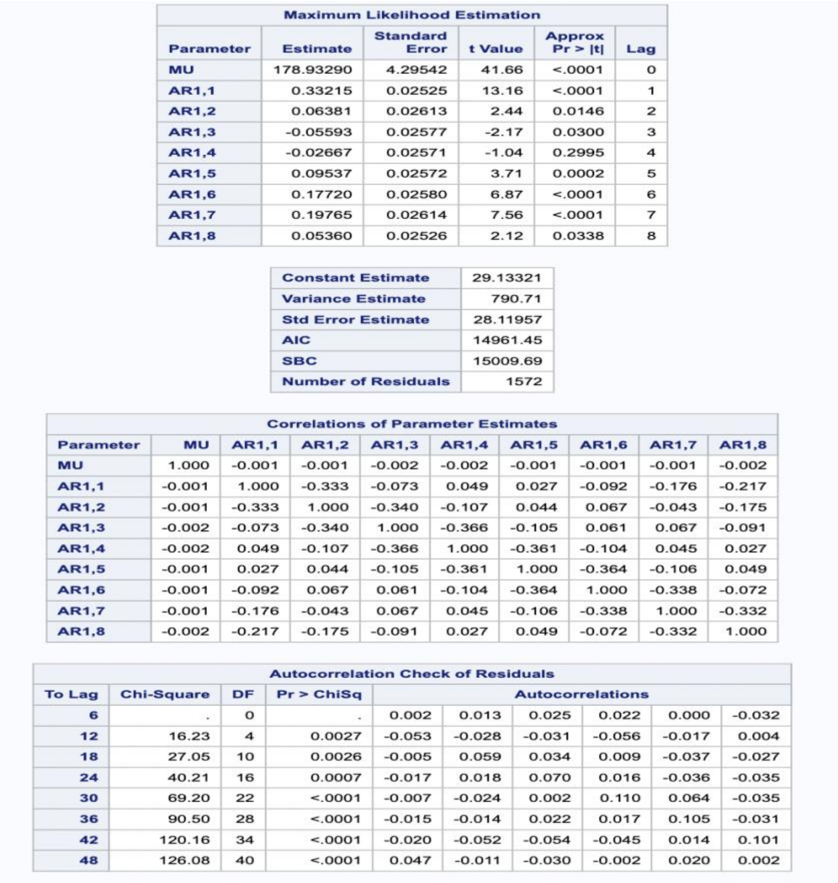
The residual plots are:





The ACF and PACF are indicating that the residuals are not random by virtue of the significant autocorrelations. In the end, we conclude that a simple trend model does not fully capture the systematic patterns over time in the global temperature series.

After a lot of trials, we fit the series to lag term 8 and perform model diagnostics checks.



Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	178.93290	4.29542	41.66	<.0001	0
AR1,1	0.33215	0.02525	13.16	<.0001	1
AR1,2	0.06381	0.02613	2.44	0.0146	2
AR1,3	-0.05593	0.02577	-2.17	0.0300	3
AR1,4	-0.02667	0.02571	-1.04	0.2995	4
AR1,5	0.09537	0.02572	3.71	0.0002	5
AR1,6	0.17720	0.02580	6.87	<.0001	6
AR1,7	0.19765	0.02614	7.56	<.0001	7
AR1,8	0.05360	0.02526	2.12	0.0338	8

Constant Estimate	29.13321
Variance Estimate	790.71
Std Error Estimate	28.11957
AIC	14961.45
SBC	15009.69
Number of Residuals	1572

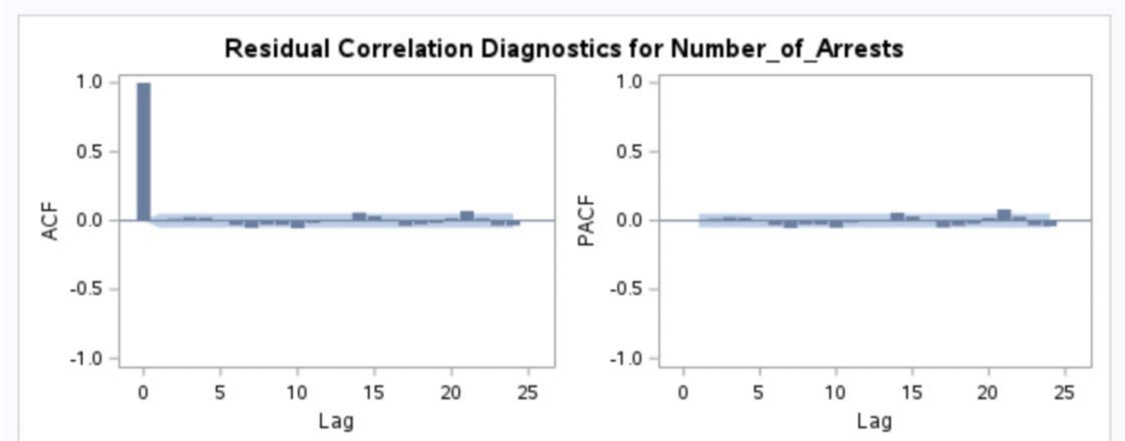
Correlations of Parameter Estimates

Parameter	MU	AR1,1	AR1,2	AR1,3	AR1,4	AR1,5	AR1,6	AR1,7	AR1,8
MU	1.000	-0.001	-0.001	-0.002	-0.002	-0.001	-0.001	-0.001	-0.002
AR1,1	-0.001	1.000	-0.333	-0.073	0.049	0.027	-0.092	-0.176	-0.217
AR1,2	-0.001	-0.333	1.000	-0.340	-0.107	0.044	0.067	-0.043	-0.175
AR1,3	-0.002	-0.073	-0.340	1.000	-0.366	-0.105	0.061	0.067	-0.091
AR1,4	-0.002	0.049	-0.107	-0.366	1.000	-0.361	-0.104	0.045	0.027
AR1,5	-0.001	0.027	0.044	-0.105	-0.361	1.000	-0.364	-0.106	0.049
AR1,6	-0.001	-0.092	0.067	0.061	-0.104	-0.364	1.000	-0.338	-0.072
AR1,7	-0.001	-0.176	-0.043	0.067	0.045	-0.106	-0.338	1.000	-0.332
AR1,8	-0.002	-0.217	-0.175	-0.091	0.027	0.049	-0.072	-0.332	1.000

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	0.002	0.013	0.025	0.022	0.000	-0.032
12	16.23	4	0.0027	-0.053	-0.028	-0.031	-0.056	-0.017	0.004
18	27.05	10	0.0026	-0.005	0.059	0.034	0.009	-0.037	-0.027
24	40.21	16	0.0007	-0.017	0.018	0.070	0.016	-0.036	-0.035
30	69.20	22	<.0001	-0.007	-0.024	0.002	0.110	0.064	-0.035
36	90.50	28	<.0001	-0.015	-0.014	0.022	0.017	0.105	-0.031
42	120.16	34	<.0001	-0.020	-0.052	-0.054	-0.045	0.014	0.101
48	126.08	40	<.0001	0.047	-0.011	-0.030	-0.002	0.020	0.002

The residual plots are:



$$\widehat{Y}_t = 29.1332 + 0.33215Y_{t-1} - 0.06381Y_{t-2} - 0.05593Y_{t-3} - 0.02663Y_{t-4} + 0.09537Y_{t-5} + 0.1772Y_{t-6} + 0.19765Y_{t-7} + 0.05360Y_{t-8}$$

With the above model, we can predict the number of arrests that may take place in Los Angeles city.

Conclusions and Recommendations from each research:

1. Given the significant age difference between male and female arrestees where female arrestees are younger than male arrestees, it is important for police officers to create programs and support systems for each gender. Targeted programs focusing on youth diversion and support for female arrestees may help mitigate risk factors associated with younger age and improve outcomes in the criminal justice system.

2. Given the significant relationship between the race and charge type of arrestees, police officers can focus on people of specific race to predict and prevent them from doing certain crimes.

3. Given the age and area, the odds of a female being arrested as a felony is 0.523 times less likely than a male, meaning a female arrested for committing felony offense is low compared to male. Also, given the gender and age, the odds of a person from 77th Street being arrested as a felony is 0.592 times less likely than the odds of a person from Southwest, in other words a person arrested in 77th street have less chances of committing felony offence compared to that of Southwest area.

4. The following model have been built to predict the daily number of arrests that may take place in Los Angeles city.

$$\begin{aligned}\hat{Y}_t = 29.1332 + 0.33215Y_{t-1} - 0.06381Y_{t-2} - 0.05593Y_{t-3} - 0.02663Y_{t-4} \\ + 0.09537Y_{t-5} + 0.1772Y_{t-6} + 0.19765Y_{t-7} + 0.05360Y_{t-8}\end{aligned}$$