Report

Prepared by Jawaria khan
Section 6B
Seat no 19122187
Course Instructor Miss Uzma
Course Data Mining
December 28, 2023

Pakistan Crimes Dataset (2012-2017)

Objective

The objective of this report is to comprehensively analyze the crime dataset of Pakistan from 2012 to 2017. By employing K-means clustering, the aim is to identify distinct crime patterns among provinces. Additionally, utilizing the time series algorithm, the report seeks to forecast future crime rates based on historical data.

Introduction and Background of the problem

K-means clustering, a powerful unsupervised learning algorithm, is utilized to segment provinces based on crime statistics, unveiling similarities and disparities in crime occurrences across different regions. This method aids in identifying clusters of provinces sharing analogous crime profiles. In conjunction, the application of time series algorithms provides a means to forecast future crime rates by examining historical data patterns, thereby empowering law enforcement agencies and policymakers with proactive strategies for crime prevention. This comprehensive approach blends unsupervised learning with predictive analytics to derive invaluable insights crucial for effective law enforcement and policymaking.

Data Collection

The crime dataset utilized for this analysis was sourced from Kaggle,This specific dataset covers crime incidents recorded in Pakistan from 2012 to 2017. The information within includes various offenses categorized by type, their occurrences per year, and the distribution across different provinces in Pakistan.
Github link: [Jaweria1234/PakistanCrimeData (github.com)]

Data Preprocessing

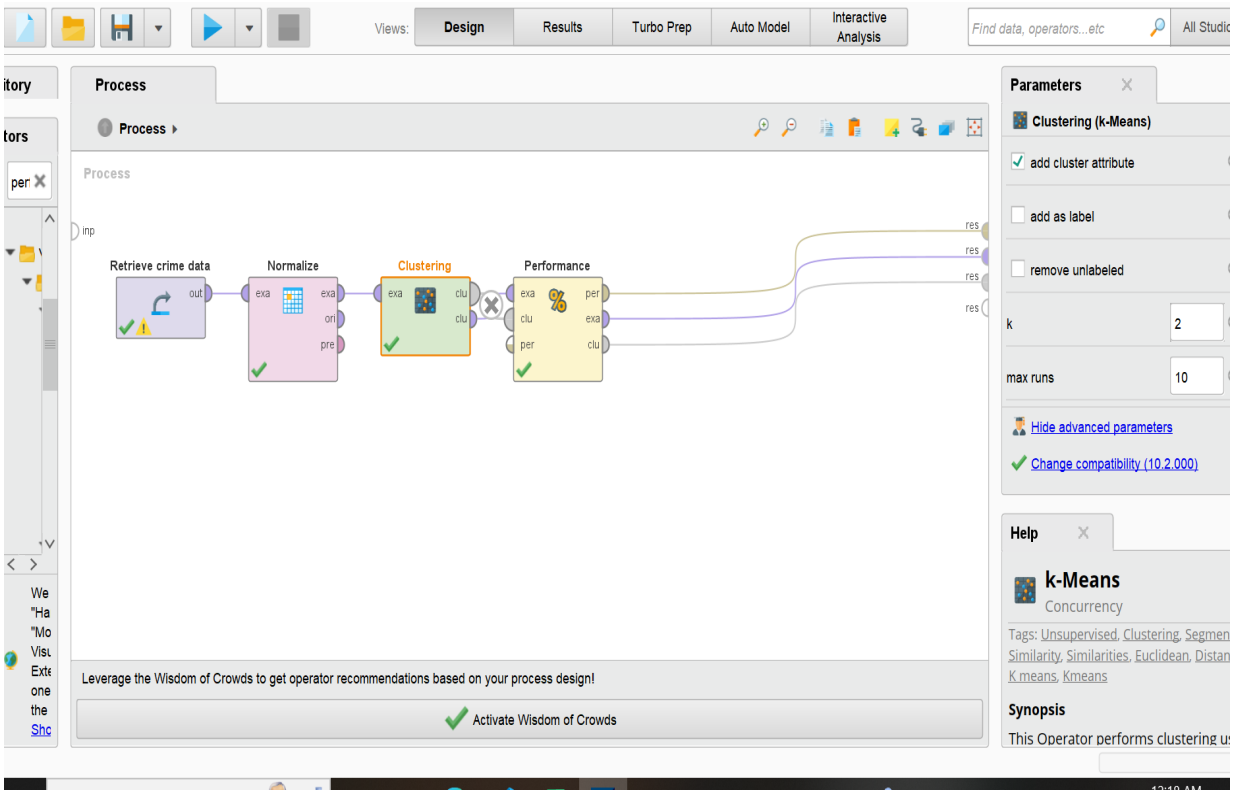
Import and clean the dataset, handle missing values if any, and encode categorical variables.

Scale the selected features if required to bring them to a similar range. This step ensures that each feature contributes equally to the clustering process.

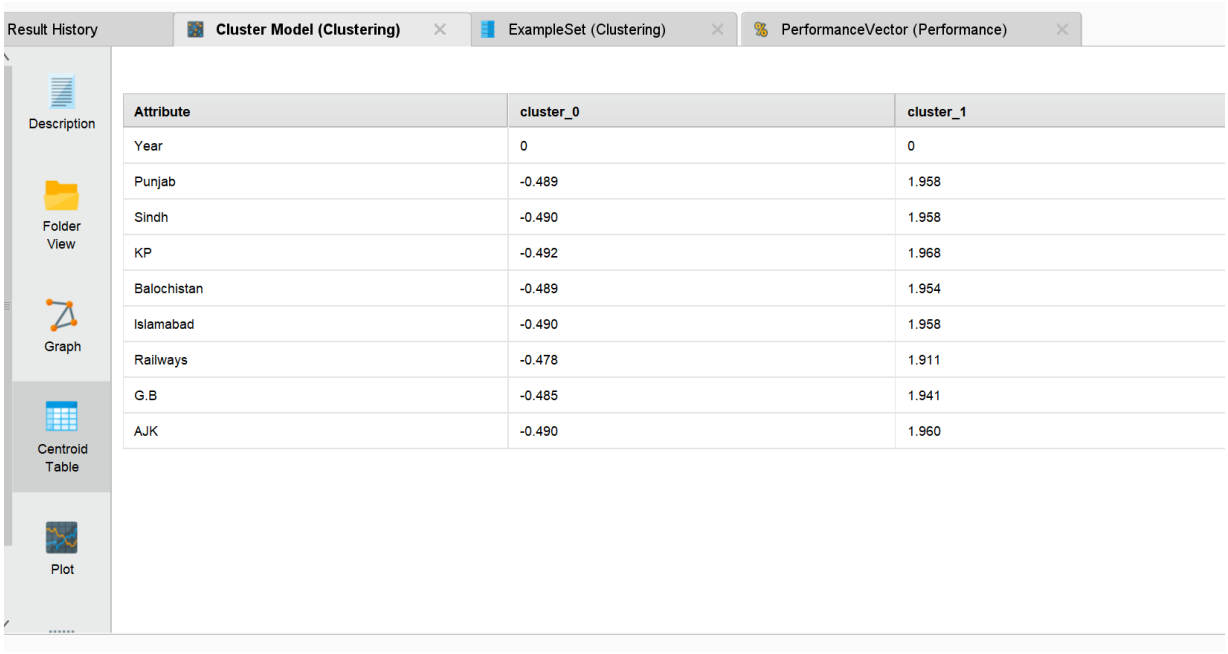
Modelling & Evaluation (K-Means Clustering)

Un-Supervised Learning Algorithms:Apply K-Means Clustering to cluster provinces based on the number and types of crimes reported over the years.

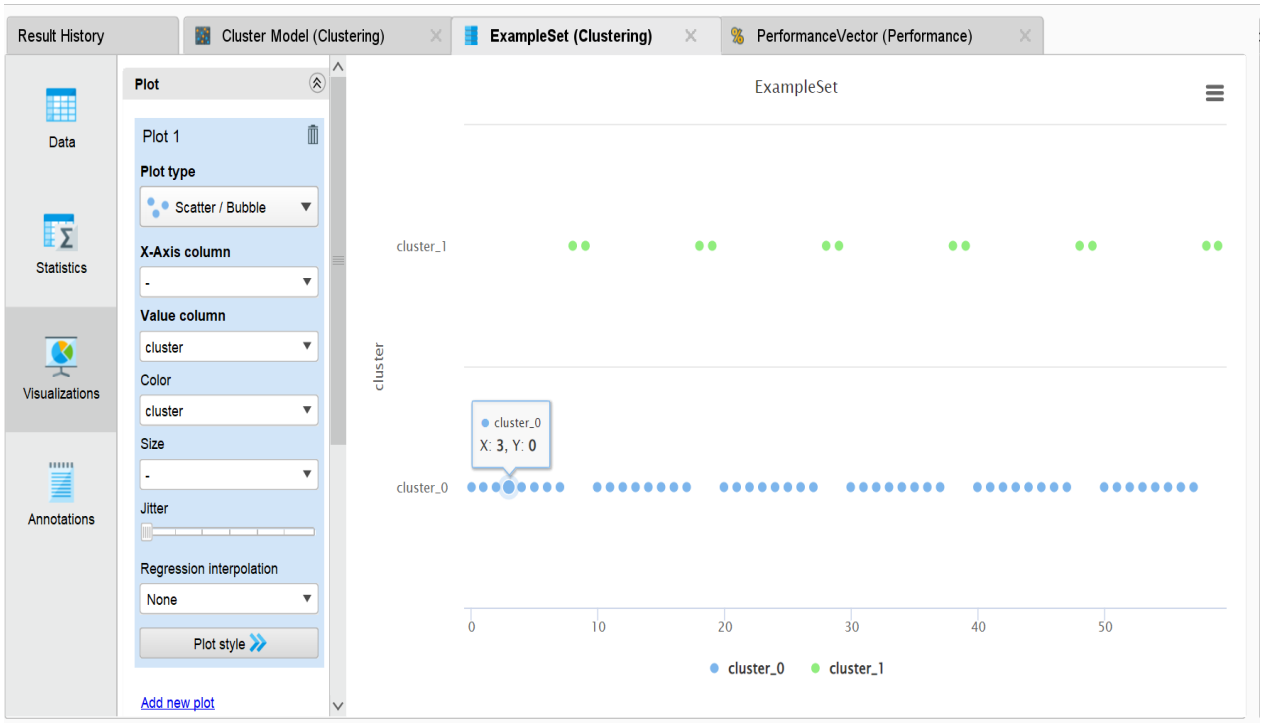
Refine the analysis by adjusting features or experimenting with different K values to improve cluster quality if needed.

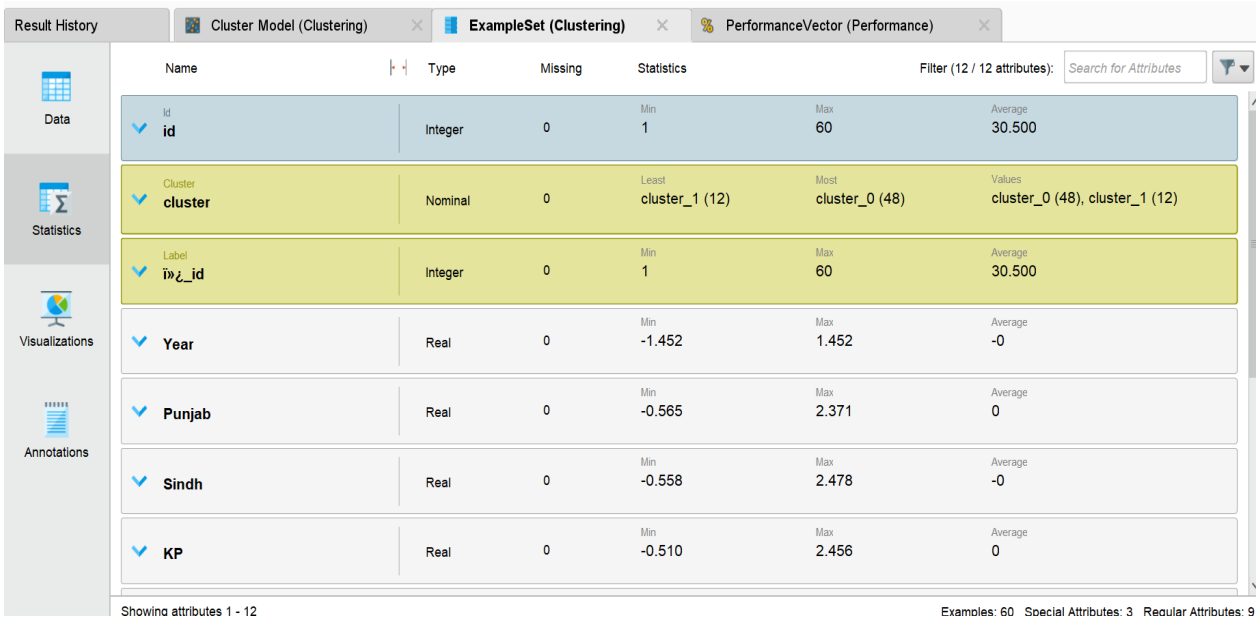


3

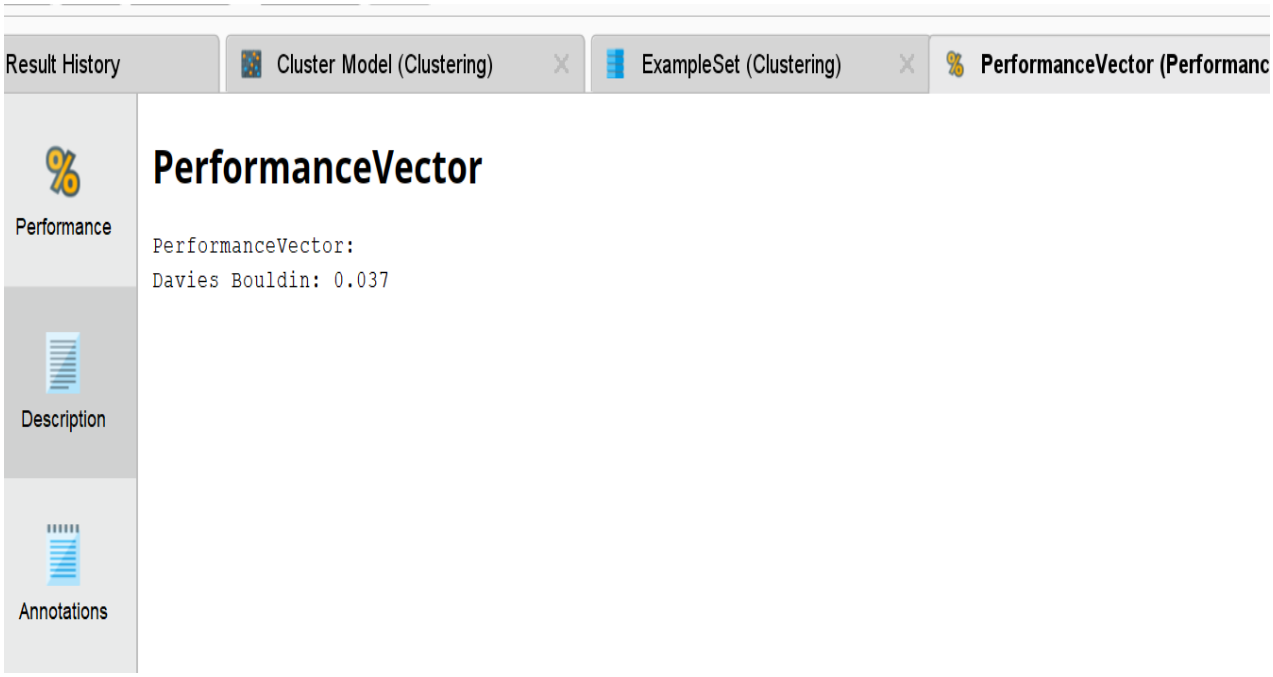


Attribute	cluster_0	cluster_1
Year	0	0
Punjab	-0.489	1.958
Sindh	-0.490	1.958
KP	-0.492	1.968
Balochistan	-0.489	1.954
Islamabad	-0.490	1.958
Railways	-0.478	1.911
G.B	-0.485	1.941
AJK	-0.490	1.960




Result History
Cluster Model (Clustering)
ExampleSet (Clustering)
PerformanceVector (Performance)
Data
Statistics
Visualizations
Annotations
Name
Type
Missing
Statistics
Filter (12 / 12 attributes):
Search for Attributes
id
Integer
0
Min 1
Max 60
Average 30.500
cluster
Nominal
0
Least cluster_1 (12)
Most cluster_0 (48)
Values cluster_0 (48), cluster_1 (12)
ï»¿_id
Integer
0
Min 1
Max 60
Average 30.500
Year
Real
0
Min -1.452
Max 1.452
Average -0
Punjab
Real
0
Min -0.565
Max 2.371
Average 0
Sindh
Real
0
Min -0.558
Max 2.478
Average -0
KP
Real
0
Min -0.510
Max 2.456
Average 0
Showing attributes 1 - 12
Examples: 60 Special Attributes: 3 Regular Attributes: 9

Results


Result History
Cluster Model (Clustering)
ExampleSet (Clustering)
PerformanceVector (Performanc
Performance
Description
Annotations
PerformanceVector
PerformanceVector:
Davies Bouldin: 0.037

Conclusions

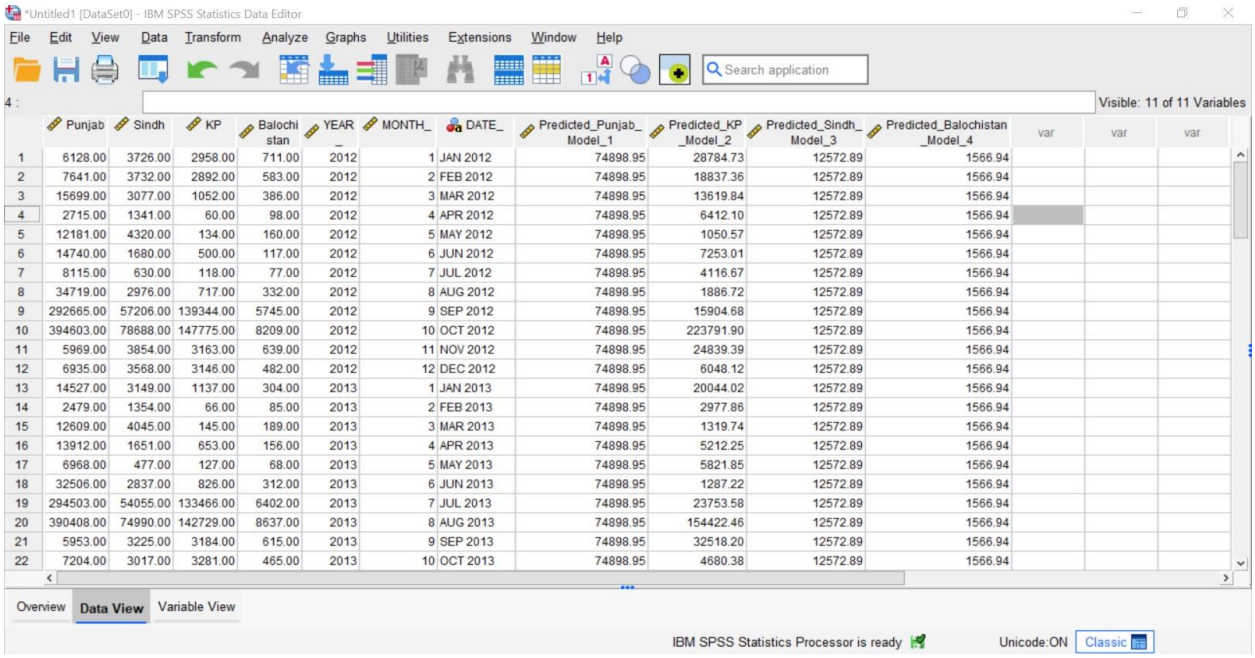
A Davies-Bouldin Index (DBI) value of **0.037** obtained from applying the K-means algorithm with two clusters suggests good separation and compactness of clusters within the dataset. The lower the DBI, the better the clustering. In this case, a value of 0.037 indicates relatively well-separated and distinct clusters, indicating a favorable performance of the K-means algorithm. This low value signifies that the clusters are -cohesive internally while being distinct from each other, which is a positive outcome for the analysis.

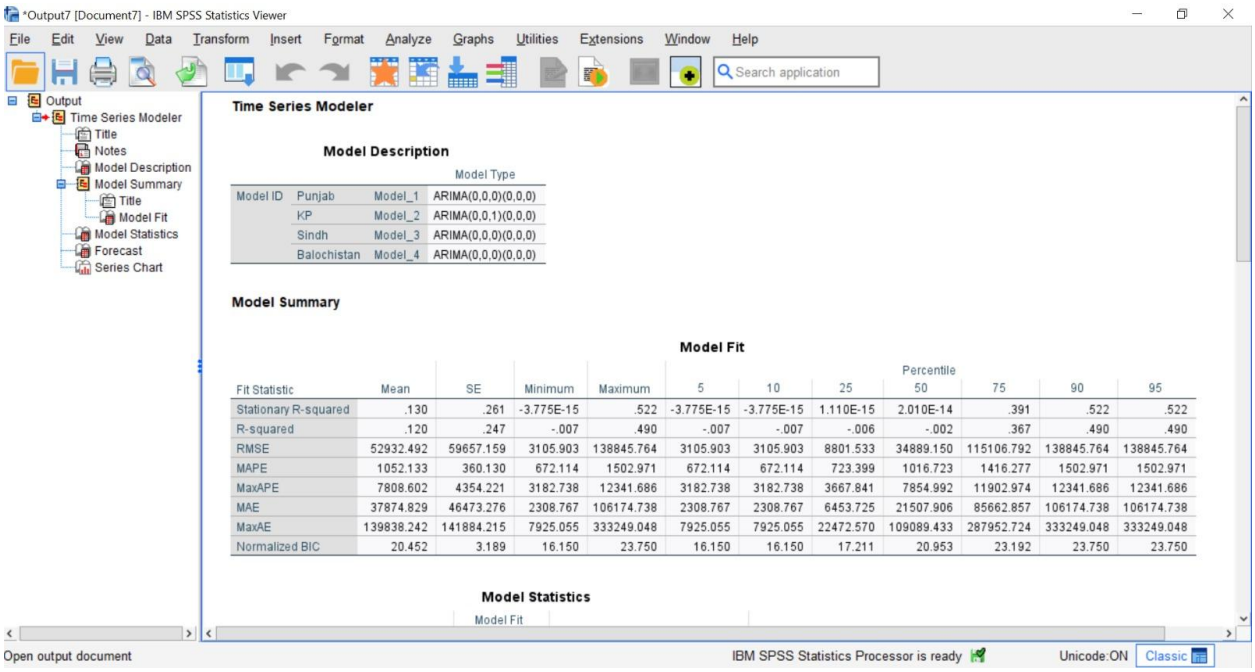
Modelling & Evaluation

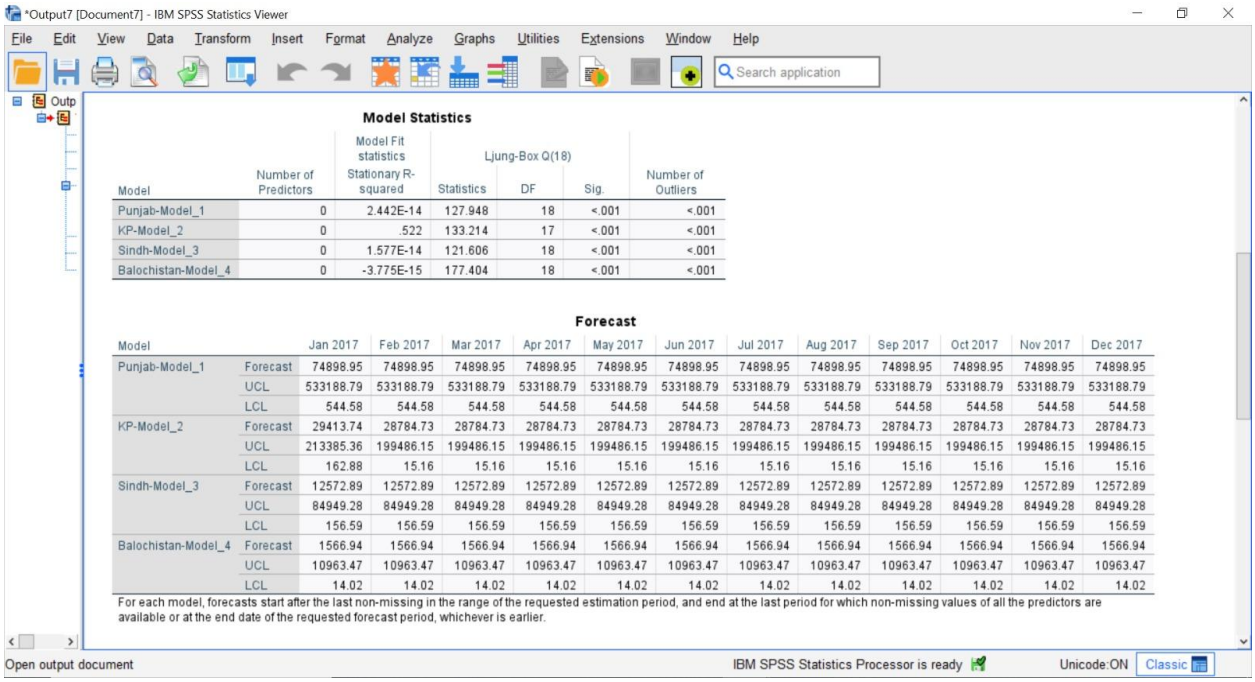
Time Series

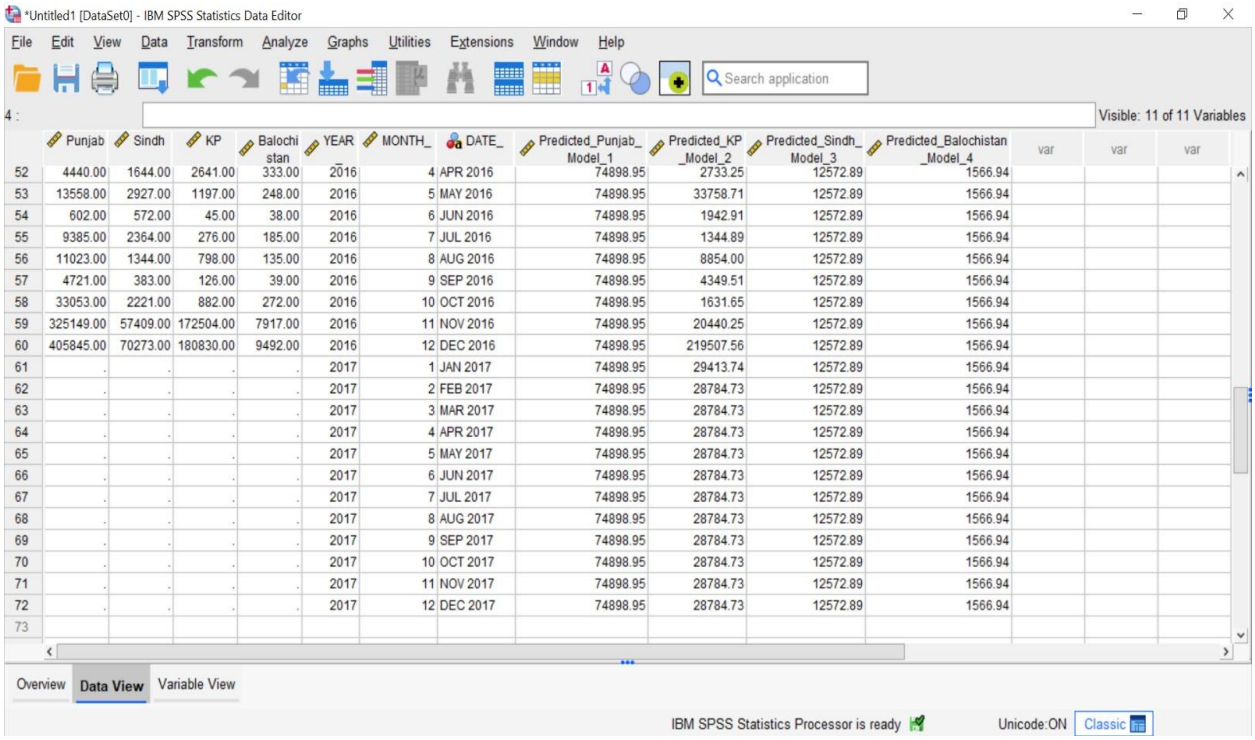
To predict the future crime rates or trends for specific offenses or provinces over time.

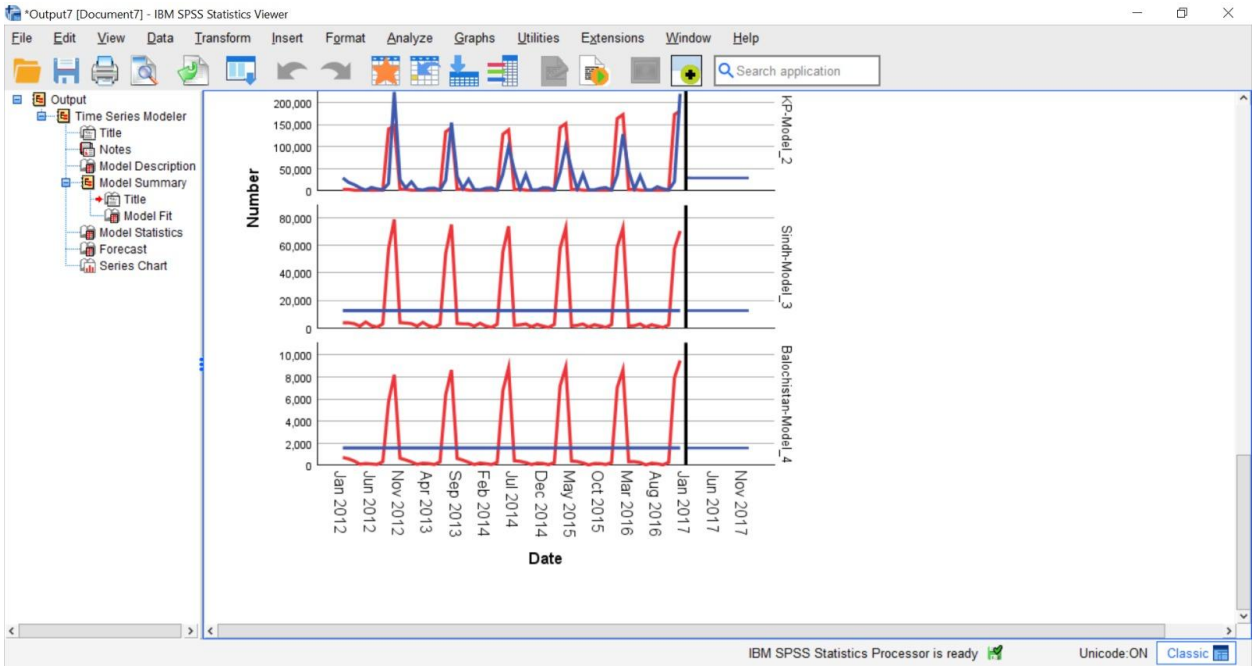
ARIMA, or AutoRegressive Integrated Moving Average, is a powerful time series forecasting method used to predict future trends based on historical data. It combines information from the past observations to forecast future values. In the context of the crime dataset, ARIMA can be utilized to predict future crime rates for specific offenses or provinces over time. The model considers the time series nature of the data and captures trends, seasonality, and patterns, enabling accurate predictions for crime rates.


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Search application
Visible: 11 of 11 Variables
Punjab Sindh KP Balochistan YEAR_ MONTH_ DATE_ Predicted_Punjab_Model_1 Predicted_KP_Model_2 Predicted_Sindh_Model_3 Predicted_Balochistan_Model_4 var var var
1 6128.00 3726.00 2958.00 711.00 2012 1 JAN 2012 74898.95 28784.73 12572.89 1566.94
2 7641.00 3732.00 2892.00 583.00 2012 2 FEB 2012 74898.95 18837.36 12572.89 1566.94
3 15699.00 3077.00 1052.00 386.00 2012 3 MAR 2012 74898.95 13619.84 12572.89 1566.94
4 2715.00 1341.00 60.00 98.00 2012 4 APR 2012 74898.95 6412.10 12572.89 1566.94
5 12181.00 4320.00 134.00 160.00 2012 5 MAY 2012 74898.95 1050.57 12572.89 1566.94
6 14740.00 1680.00 500.00 117.00 2012 6 JUN 2012 74898.95 7253.01 12572.89 1566.94
7 8115.00 630.00 118.00 77.00 2012 7 JUL 2012 74898.95 4116.67 12572.89 1566.94
8 34719.00 2976.00 717.00 332.00 2012 8 AUG 2012 74898.95 1886.72 12572.89 1566.94
9 292665.00 57206.00 139344.00 5745.00 2012 9 SEP 2012 74898.95 15904.68 12572.89 1566.94
10 394603.00 78688.00 147775.00 8209.00 2012 10 OCT 2012 74898.95 223791.90 12572.89 1566.94
11 5969.00 3854.00 3163.00 639.00 2012 11 NOV 2012 74898.95 24839.39 12572.89 1566.94
12 6935.00 3568.00 3146.00 482.00 2012 12 DEC 2012 74898.95 6048.12 12572.89 1566.94
13 14527.00 3149.00 1137.00 304.00 2013 1 JAN 2013 74898.95 20044.02 12572.89 1566.94
14 2479.00 1354.00 66.00 85.00 2013 2 FEB 2013 74898.95 2977.86 12572.89 1566.94
15 12609.00 4045.00 145.00 189.00 2013 3 MAR 2013 74898.95 1319.74 12572.89 1566.94
16 13912.00 1651.00 653.00 156.00 2013 4 APR 2013 74898.95 5212.25 12572.89 1566.94
17 6968.00 477.00 127.00 68.00 2013 5 MAY 2013 74898.95 5821.85 12572.89 1566.94
18 32506.00 2837.00 826.00 312.00 2013 6 JUN 2013 74898.95 1287.22 12572.89 1566.94
19 294503.00 54055.00 133466.00 6402.00 2013 7 JUL 2013 74898.95 23753.58 12572.89 1566.94
20 390408.00 74990.00 142729.00 8637.00 2013 8 AUG 2013 74898.95 154422.46 12572.89 1566.94
21 5953.00 3225.00 3184.00 615.00 2013 9 SEP 2013 74898.95 32518.20 12572.89 1566.94
22 7204.00 3017.00 3281.00 465.00 2013 10 OCT 2013 74898.95 4680.38 12572.89 1566.94
Overview Data View Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON Classic


*Output7 [Document7] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help
Search application
Output
Time Series Modeler
Title
Notes
Model Description
Model Summary
Title
Model Fit
Model Statistics
Forecast
Series Chart
Time Series Modeler
Model Description
Model Type
Model ID Punjab Model_1 ARIMA(0,0,0)(0,0,0)
KP Model_2 ARIMA(0,0,1)(0,0,0)
Sindh Model_3 ARIMA(0,0,0)(0,0,0)
Balochistan Model_4 ARIMA(0,0,0)(0,0,0)
Model Summary
Model Fit
Percentile
Fit Statistic Mean SE Minimum Maximum 5 10 25 50 75 90 95
Stationary R-squared .130 .261 -3.775E-15 .522 -3.775E-15 -3.775E-15 1.110E-15 2.010E-14 .391 .522 .522
R-squared .120 .247 -.007 .490 -.007 -.007 -.006 -.002 .367 .490 .490
RMSE 52932.492 59657.159 3105.903 138845.764 3105.903 3105.903 8801.533 34889.150 115106.792 138845.764 138845.764
MAPE 1052.133 360.130 672.114 1502.971 672.114 672.114 723.399 1016.723 1416.277 1502.971 1502.971
MaxAPE 7808.602 4354.221 3182.738 12341.686 3182.738 3182.738 3667.841 7854.992 11902.974 12341.686 12341.686
MAE 37874.829 46473.276 2308.767 106174.738 2308.767 2308.767 6453.725 21507.906 85662.857 106174.738 106174.738
MaxAE 139838.242 141884.215 7925.055 333249.048 7925.055 7925.055 22472.570 109089.433 287952.724 333249.048 333249.048
Normalized BIC 20.452 3.189 16.150 23.750 16.150 16.150 17.211 20.953 23.192 23.750 23.750
Model Statistics
Model Fit
Open output document
IBM SPSS Statistics Processor is ready
Unicode:ON Classic


*Output7 [Document7] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help
Search application
Outp
Model Statistics
Model | Number of Predictors | Model Fit statistics: Stationary R-squared | Ljung-Box Q(18): Statistics | DF | Sig. | Number of Outliers
Punjab-Model_1 | 0 | 2.442E-14 | 127.948 | 18 | <.001 | <.001
KP-Model_2 | 0 | .522 | 133.214 | 17 | <.001 | <.001
Sindh-Model_3 | 0 | 1.577E-14 | 121.606 | 18 | <.001 | <.001
Balochistan-Model_4 | 0 | -3.775E-15 | 177.404 | 18 | <.001 | <.001
Forecast
Model | | Jan 2017 | Feb 2017 | Mar 2017 | Apr 2017 | May 2017 | Jun 2017 | Jul 2017 | Aug 2017 | Sep 2017 | Oct 2017 | Nov 2017 | Dec 2017
Punjab-Model_1 | Forecast | 74898.95 | 74898.95 | 74898.95 | 74898.95 | 74898.95 | 74898.95 | 74898.95 | 74898.95 | 74898.95 | 74898.95 | 74898.95 | 74898.95
UCL | 533188.79 | 533188.79 | 533188.79 | 533188.79 | 533188.79 | 533188.79 | 533188.79 | 533188.79 | 533188.79 | 533188.79 | 533188.79 | 533188.79
LCL | 544.58 | 544.58 | 544.58 | 544.58 | 544.58 | 544.58 | 544.58 | 544.58 | 544.58 | 544.58 | 544.58 | 544.58
KP-Model_2 | Forecast | 29413.74 | 28784.73 | 28784.73 | 28784.73 | 28784.73 | 28784.73 | 28784.73 | 28784.73 | 28784.73 | 28784.73 | 28784.73 | 28784.73
UCL | 213385.36 | 199486.15 | 199486.15 | 199486.15 | 199486.15 | 199486.15 | 199486.15 | 199486.15 | 199486.15 | 199486.15 | 199486.15 | 199486.15
LCL | 162.88 | 15.16 | 15.16 | 15.16 | 15.16 | 15.16 | 15.16 | 15.16 | 15.16 | 15.16 | 15.16 | 15.16
Sindh-Model_3 | Forecast | 12572.89 | 12572.89 | 12572.89 | 12572.89 | 12572.89 | 12572.89 | 12572.89 | 12572.89 | 12572.89 | 12572.89 | 12572.89 | 12572.89
UCL | 84949.28 | 84949.28 | 84949.28 | 84949.28 | 84949.28 | 84949.28 | 84949.28 | 84949.28 | 84949.28 | 84949.28 | 84949.28 | 84949.28
LCL | 156.59 | 156.59 | 156.59 | 156.59 | 156.59 | 156.59 | 156.59 | 156.59 | 156.59 | 156.59 | 156.59 | 156.59
Balochistan-Model_4 | Forecast | 1566.94 | 1566.94 | 1566.94 | 1566.94 | 1566.94 | 1566.94 | 1566.94 | 1566.94 | 1566.94 | 1566.94 | 1566.94 | 1566.94
UCL | 10963.47 | 10963.47 | 10963.47 | 10963.47 | 10963.47 | 10963.47 | 10963.47 | 10963.47 | 10963.47 | 10963.47 | 10963.47 | 10963.47
LCL | 14.02 | 14.02 | 14.02 | 14.02 | 14.02 | 14.02 | 14.02 | 14.02 | 14.02 | 14.02 | 14.02 | 14.02
For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.
Open output document
IBM SPSS Statistics Processor is ready
Unicode:ON
Classic


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Search application
4 :
Visible: 11 of 11 Variables
Punjab | Sindh | KP | Balochistan | YEAR_ | MONTH_ | DATE_ | Predicted_Punjab_Model_1 | Predicted_KP_Model_2 | Predicted_Sindh_Model_3 | Predicted_Balochistan_Model_4 | var | var | var
52 | 4440.00 | 1644.00 | 2641.00 | 333.00 | 2016 | 4 | APR 2016 | 74898.95 | 2733.25 | 12572.89 | 1566.94
53 | 13558.00 | 2927.00 | 1197.00 | 248.00 | 2016 | 5 | MAY 2016 | 74898.95 | 33758.71 | 12572.89 | 1566.94
54 | 602.00 | 572.00 | 45.00 | 38.00 | 2016 | 6 | JUN 2016 | 74898.95 | 1942.91 | 12572.89 | 1566.94
55 | 9385.00 | 2364.00 | 276.00 | 185.00 | 2016 | 7 | JUL 2016 | 74898.95 | 1344.89 | 12572.89 | 1566.94
56 | 11023.00 | 1344.00 | 798.00 | 135.00 | 2016 | 8 | AUG 2016 | 74898.95 | 8854.00 | 12572.89 | 1566.94
57 | 4721.00 | 383.00 | 126.00 | 39.00 | 2016 | 9 | SEP 2016 | 74898.95 | 4349.51 | 12572.89 | 1566.94
58 | 33053.00 | 2221.00 | 882.00 | 272.00 | 2016 | 10 | OCT 2016 | 74898.95 | 1631.65 | 12572.89 | 1566.94
59 | 325149.00 | 57409.00 | 172504.00 | 7917.00 | 2016 | 11 | NOV 2016 | 74898.95 | 20440.25 | 12572.89 | 1566.94
60 | 405845.00 | 70273.00 | 180830.00 | 9492.00 | 2016 | 12 | DEC 2016 | 74898.95 | 219507.56 | 12572.89 | 1566.94
61 | . | . | . | . | 2017 | 1 | JAN 2017 | 74898.95 | 29413.74 | 12572.89 | 1566.94
62 | . | . | . | . | 2017 | 2 | FEB 2017 | 74898.95 | 28784.73 | 12572.89 | 1566.94
63 | . | . | . | . | 2017 | 3 | MAR 2017 | 74898.95 | 28784.73 | 12572.89 | 1566.94
64 | . | . | . | . | 2017 | 4 | APR 2017 | 74898.95 | 28784.73 | 12572.89 | 1566.94
65 | . | . | . | . | 2017 | 5 | MAY 2017 | 74898.95 | 28784.73 | 12572.89 | 1566.94
66 | . | . | . | . | 2017 | 6 | JUN 2017 | 74898.95 | 28784.73 | 12572.89 | 1566.94
67 | . | . | . | . | 2017 | 7 | JUL 2017 | 74898.95 | 28784.73 | 12572.89 | 1566.94
68 | . | . | . | . | 2017 | 8 | AUG 2017 | 74898.95 | 28784.73 | 12572.89 | 1566.94
69 | . | . | . | . | 2017 | 9 | SEP 2017 | 74898.95 | 28784.73 | 12572.89 | 1566.94
70 | . | . | . | . | 2017 | 10 | OCT 2017 | 74898.95 | 28784.73 | 12572.89 | 1566.94
71 | . | . | . | . | 2017 | 11 | NOV 2017 | 74898.95 | 28784.73 | 12572.89 | 1566.94
72 | . | . | . | . | 2017 | 12 | DEC 2017 | 74898.95 | 28784.73 | 12572.89 | 1566.94
73
Overview
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON
Classic



Results & Conclusions

In the results and conclusions, the ARIMA model's performance, accuracy, and insights derived from the forecasts will provide valuable information for understanding and potentially mitigating future crime rates in specific regions or for particular offenses.