Spatial and Temporal Dynamics of Dengue Healthcare Data

(Course Project for Advance Techniques in Data Science)
Aqib Javaid 2024-MS-DS-129
Submitted to: Prof. Dr. Muhammad Awais Hassan

1 Problem Statement and Motivation

Dengue fever is a significant public health concern, with its transmission being influenced by spatial and temporal factors. The rapid spread of dengue cases challenges healthcare systems, particularly in identifying and managing hotspots efficiently. This study aims to analyze the spatial and temporal dynamics of dengue cases using healthcare data to provide actionable insights for resource allocation and epidemic control.

1.1 Significance

- Spatial Dynamics: Understanding geographical hotspots of dengue cases can aid in targeting interventions such as fumigation, awareness campaigns, and healthcare resource allocation.
- Temporal Trends: Analyzing time-based patterns helps in predicting peak seasons, enabling proactive healthcare measures.
- Gender and Age Analysis: Insights into demographic distributions can improve personalized healthcare strategies.
- Complaints Analysis: Studying the primary symptoms reported can help in early diagnosis and tailored treatment plans.

1.2 Objectives

- Data Integration and Preprocessing: Clean, standardize, and integrate multiple datasets containing dengue-related complaints, demographics, and geographical information.
- Spatial Analysis: Identify geographical hotspots using clustering techniques and visualize them on heatmaps.
- Temporal Analysis: Perform time-series analysis to uncover trends, seasonality, and anomalies in dengue case reports.
- Demographic Study: Analyze the distribution of dengue cases across age groups and genders.
- Predictive Modeling: Use machine learning models like Support Vector Machines (SVM) to estimate the intensity of dengue cases in different regions.
- Time Series Forecasting: Forecasting time series trend of total males and females using Machine Learning algorithms.

1.3 Potential Impact

- Healthcare Preparedness: Equip authorities with tools to predict and prepare for dengue outbreaks effectively.
- Targeted Interventions: Enable focused efforts in high-risk regions, optimizing resource usage.
- Policy Making: Provide data-driven insights to shape policies for vector control and healthcare improvements.
- Community Awareness: Enhance public awareness by highlighting high-risk areas and periods.

2 Questions for Analysis

2.1 Spatial Analysis Questions

1. Where are the key geographical hotspots of dengue cases during the studied period?
2. How do dengue case distributions vary across different locations (urban vs. rural)?
3. How effectively can hotspots of existing case densities be visualized using SVM-based heatmaps?
4. Which regions consistently exhibit high or low dengue case densities?

2.2 Temporal Analysis Questions

1. Are there distinct seasonal patterns in dengue cases across the studied months (May-August)?
2. What are the peak periods for dengue cases during this timeframe?
3. Can anomalies in case trends (e.g., sudden spikes) provide early warning signals for potential outbreaks?

2.3 Demographic Analysis Questions

1. Which age groups are most affected by dengue during this time?
2. Is there a significant gender difference in dengue incidence?
3. How do complaint patterns vary by age group and gender?

2.4 Symptom Analysis Questions

1. What are the most commonly reported symptoms among dengue patients?
2. How do symptom trends differ across regions or demographics (e.g., younger vs. older patients)?

2.5 Predictive Analysis Questions

1. How effectively do SVM-based heatmaps represent the spatial distribution of dengue case hotspots?
2. Can K-Means clustering accurately aggregate dengue case data based on geographical coordinates and case frequencies?
3. What are the key limitations of Random Forest Regressor and SVM in forecasting time-series data with only the Date feature?
4. How can the prediction of total male and female patients, along with their age groups, be improved given the limited feature set?
5. What adjustments or additional features could enhance the R^2 score and overall accuracy of time-series forecasting models?

3 Data Collection

The dataset used for this analysis was acquired from one of my professors and contains records of dengue patients registered at Mayo Hospital during the year 2011. This year was chosen as it marked a peak in dengue cases, with a significant surge and a large wave of patients.

The dataset is already privatized and complies with HIPAA (Health Insurance Portability and Accountability Act) regulations, ensuring the privacy and security of sensitive information. Acquiring a more recent dataset would require formal meetings with hospital authorities and considerable time, making this dataset an optimal choice for the study.

This data is highly relevant to the problem as it captures the spatial, temporal, and demographic dynamics of dengue cases during a critical period, providing sufficient features and observations for meaningful analysis.

4 Data Wrangling (Preprocessing)

This section outlines the steps taken to clean and preprocess the dataset to ensure it is ready for analysis. Figures and screenshots are included to illustrate the process.

4.1 Cleaning the Dataset

4.1.1 Handling Missing Values

- Identified missing values in the dataset using `isna().sum()`.
- Filled missing values in the *1st complaint*, *2nd Complaint*, and *3rd Complaint* columns with their respective mode values.
- Dropped rows with missing dates, as these are crucial for temporal analysis.
- Removed rows with missing latitude and longitude after attempting imputation based on standardized addresses. See Figure 1 and 2.

4.1.2 Address Standardization

- Standardized inconsistent address names (e.g., "Guwalmandi" to "Gowalmandi").
- Mapped addresses to corresponding latitude and longitude values.

See Figure 3.

4.2 Removing or Addressing Outliers

4.2.1 Address Frequencies

- Identified outliers in address frequencies using the IQR method.
- Removed or aggregated low-frequency addresses.

See Figure 4.

4.2.2 Complaint Frequencies

- Detected outliers in the *1st complaint*, *2nd Complaint*, and *3rd Complaint* columns using the IQR method.

See Figure 5 to 10.

```
whole_data.isna().sum()
```

```
Date                1
Age                 1
Sex                 1
Address             0
1st complaint     496
2nd Complaint     529
3rd Complaint    1259
dtype: int64
```

```
whole_data[whole_data['Date'].isna()]
```

	Date	Age	Sex	Address	1st complaint	2nd Complaint	3rd Complaint
97505	NaN	NaN	NaN	Aug-11	NaN	NaN	NaN

```
whole_data = whole_data.dropna(subset=['Date'])
whole_data
```

Figure 1: Handling Date missing value.

4.2.3 Age Outliers

- Identified age outliers using the IQR method and visualized them with a boxplot.

See Figure 11.

4.3 Converting Data Types

- Converted the *Date* column to datetime format for accurate time-series analysis.
- Converted the *Age* column to numeric format, coercing invalid entries.

See Figure 12.

4.4 Feature Engineering

Creating New Variables

- Created *Latitude* and *Longitude* columns based on standardized addresses.
- Derived *Age Group* from the *Age* column using categorical bins.
- Extracted the *Month* from the *Date* column for monthly trend analysis.
- Aggregated complaints into a single column using a melting operation.

See Figure 13.

4.4.1 Standardization and Normalization

- Normalized patient counts for hotspot visualization using MinMaxScaler.

See Figure 14.

```
whole_data['3rd Complaint']=whole_data['3rd Complaint'].fillna(mode_value)
```

```
C:\Users\drfak\AppData\Local\Temp\ipykernel_7444\1230600248.py:1: SettingWi
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs
  whole_data['3rd Complaint']=whole_data['3rd Complaint'].fillna(mode_value
```

```
whole_data.isna().sum()
```

```
Date             0
Age              0
Sex              0
Address          0
1st complaint    0
2nd Complaint    0
3rd Complaint    0
dtype: int64
```

Figure 2: Handling Complaints missing value.

4.5 Grouping and Aggregating Data

Grouped data by *Date*, *Address*, *Latitude*, and *Longitude*, and *age-groups* to calculate:

- Total male and female cases.
- Dominant complaints for each location.
- Most affected age groups. See Figures 15-24

5 Support Vector Machine (SVM) for Hotspots Generation

5.1 Introduction to SVM

Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression. For regression tasks, SVM aims to find a function $f(x)$ that deviates from the actual output y by no more than a specified margin ϵ, while minimizing the model complexity.

The objective function for SVM regression is:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\sum(\xi_i + \xi_i^*)$$

Subject to:

$$y_i - (w \cdot x_i + b) \leq \epsilon + \xi_i$$
$$(w \cdot x_i + b) - y_i \leq \epsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \geq 0$$

Where:

- w: Weight vector.

	Date	Age	Sex	1st complaint	2nd Complaint	3rd Complaint	Address	Latitude	Longitude
0	01.05.2011	18	Male	fever	bodyaches	Headache	KOT ABDUL MALIK	31.620420	74.234381
1	01.05.2011	20	Male	pain	bleeding	RTA	Gulshan-e-Ravi	31.552170	74.275290
2	01.05.2011	40	Male	Dyspnea	chest pain	RTI	Gawalmandi	31.571870	74.318260
3	01.05.2011	24	Male	headache	allergy	Vomiting	KOT ABDUL MALIK	31.620420	74.234381
6	01.05.2011	20	Male	pain	bleeding	RTA	Gawalmandi	31.571870	74.318260
...	...	...	...	...	...	...	...	...	...
83452	31.08.2011	31	Female	Fever	Chills	Body pains	Gujranwala	32.166351	74.195900
83453	31.08.2011	35	Female	Vomiting	fever	nausea	Anarkali Bazaar Lahore	31.569800	74.312000
83454	31.08.2011	35	Female	Arthritis	Anxiety	arthralgia	Gawalmandi	31.571870	74.318260
83456	31.08.2011	46	Male	Palpitation	anxiety	arthralgia	Lahori Gate	31.577410	74.313430
83457	31.08.2011	18	Female	Fever	headache	chills	Gawalmandi	31.571870	74.318260

74293 rows × 9 columns

Figure 3: Standardization of addresses and mapping to latitude and longitude.

- b: Bias term.
- ξ, ξ^*: Slack variables for deviations beyond ϵ.
- C: Regularization parameter balancing margin width and tolerance.

5.2 Radial Basis Function (RBF) Kernel

The RBF kernel is defined as:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

Where:

- γ: Controls the influence of a single training example. Smaller values imply broader influence; larger values make it more localized.

5.3 Hotspot Generation

For hotspot generation:

1. Latitude and Longitude are used as features, and normalized patient counts as the target.
2. MinMaxScaler normalizes the patient counts to $[0, 1]$.
3. SVM with RBF kernel is trained to predict patient density.
4. Predicted intensities are visualized as a heatmap using Folium.

5.4 Significance of Parameters

- γ: A high value (e.g., 10^6) makes the SVM sensitive to localized regions.
- C: Controls overfitting; $C = 1$ strikes a balance between margin width and error tolerance.

See Figure 25.

—

```
Q1: 144.0
Q2 (Median): 337.0
Q3: 883.5
IQR: 739.5
Lower Limit: -965.25
Upper Limit: 1992.75
Number of outliers: 48
```

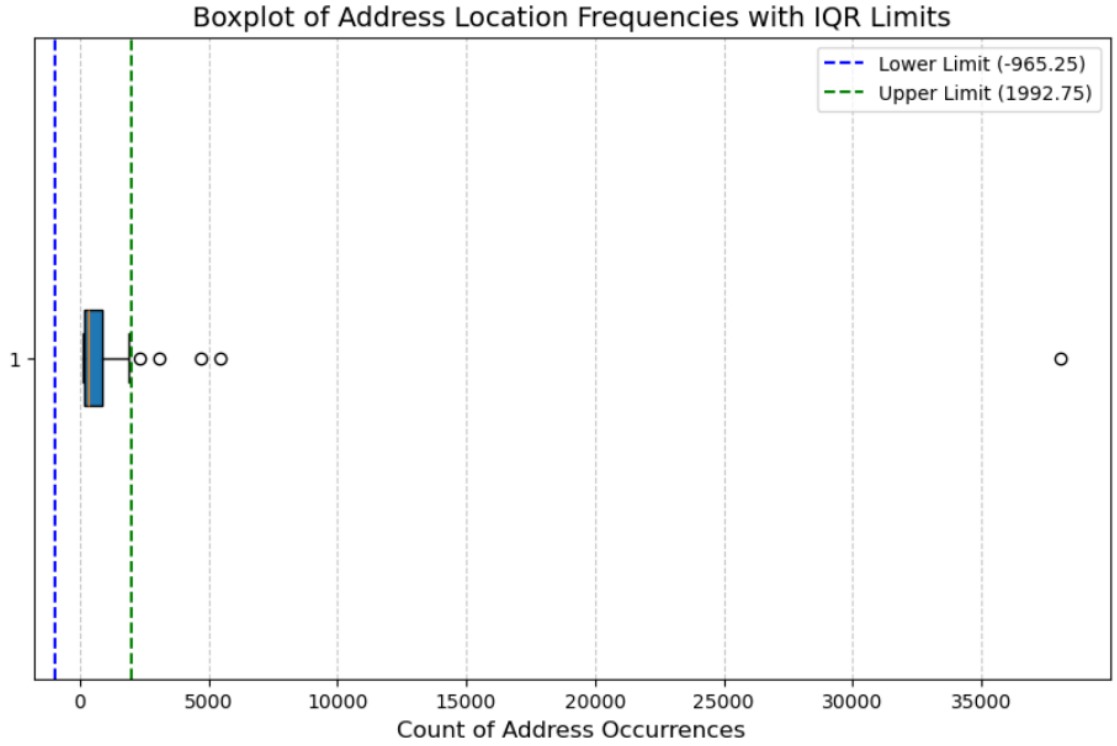

Figure 4: Boxplot showing outliers in address frequencies.

6 STL Decomposition and Anomaly Detection

6.1 STL Decomposition

STL (Seasonal-Trend Decomposition using LOESS) decomposes a time series into:

$$y_t = T_t + S_t + R_t$$

Where:

- y_t: Original time series.
- T_t: Trend component.
- S_t: Seasonal component.
- R_t: Residual component (noise).

6.1.1 Advantages of STL

- Handles non-linear seasonality.
- Robust to missing data.

	1st complaint	Count
0	fever	14085
1	chills	5036
2	headache	4442
3	vomiting	3745
4	bodyaches	3349
...	...	...
247	Muscle pain	7
248	Arthralgia	5
249	wounds	5
250	redness	5
251	pain eyes	3

252 rows × 2 columns

Figure 5: Count of 1st Complaint

```
Q1: 19.0
Q2 (Median): 43.5
Q3: 201.0
IQR: 182.0
Lower Limit: -254.0
Upper Limit: 474.0
Number of outliers: 252
```

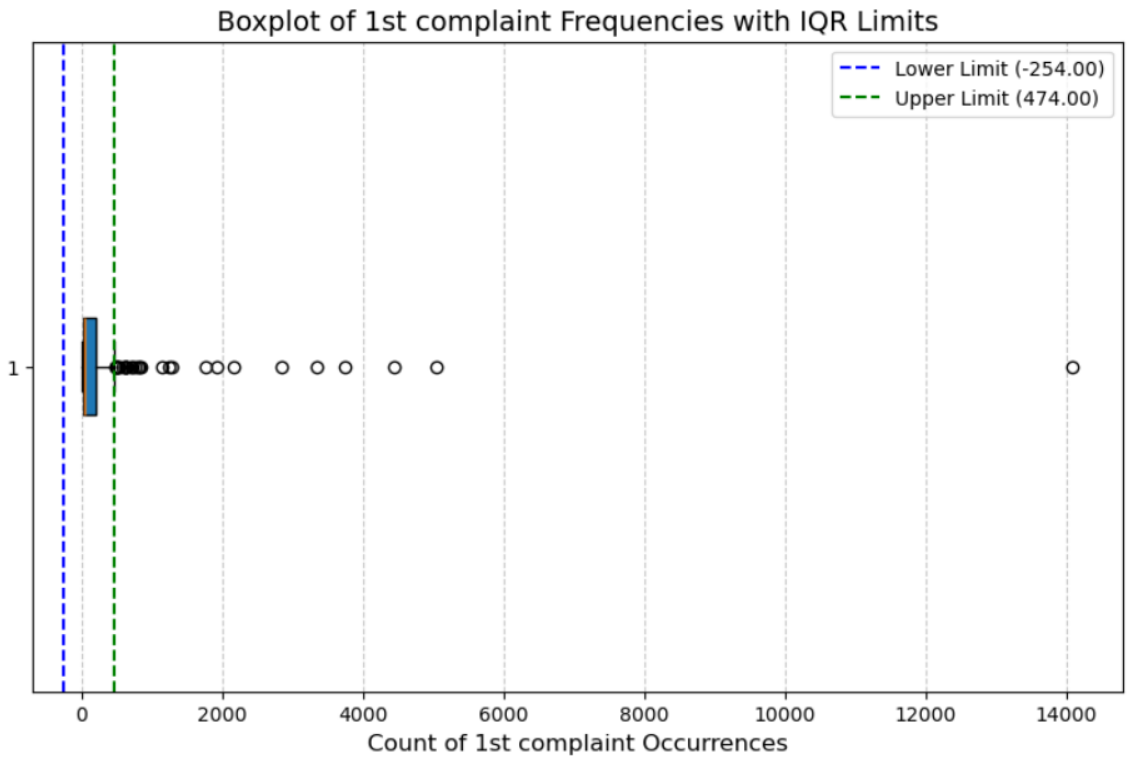

Figure 6: Boxplots of 1st complaint frequencies showing outliers.

	2nd Complaint	Count
0	fever	12034
1	chills	6433
2	headache	4072
3	vomiting	3526
4	nausea	3470
...	...	...
247	Anorexia	3
248	burning	3
249	Fbeye	2
250	Pain body	1
251	Lethargy	1

252 rows × 2 columns

Figure 7: Count of 2nd Complaint

```
Q1: 15.0
Q2 (Median): 39.0
Q3: 175.75
IQR: 160.75
Lower Limit: -226.125
Upper Limit: 416.875
Number of outliers: 252
```

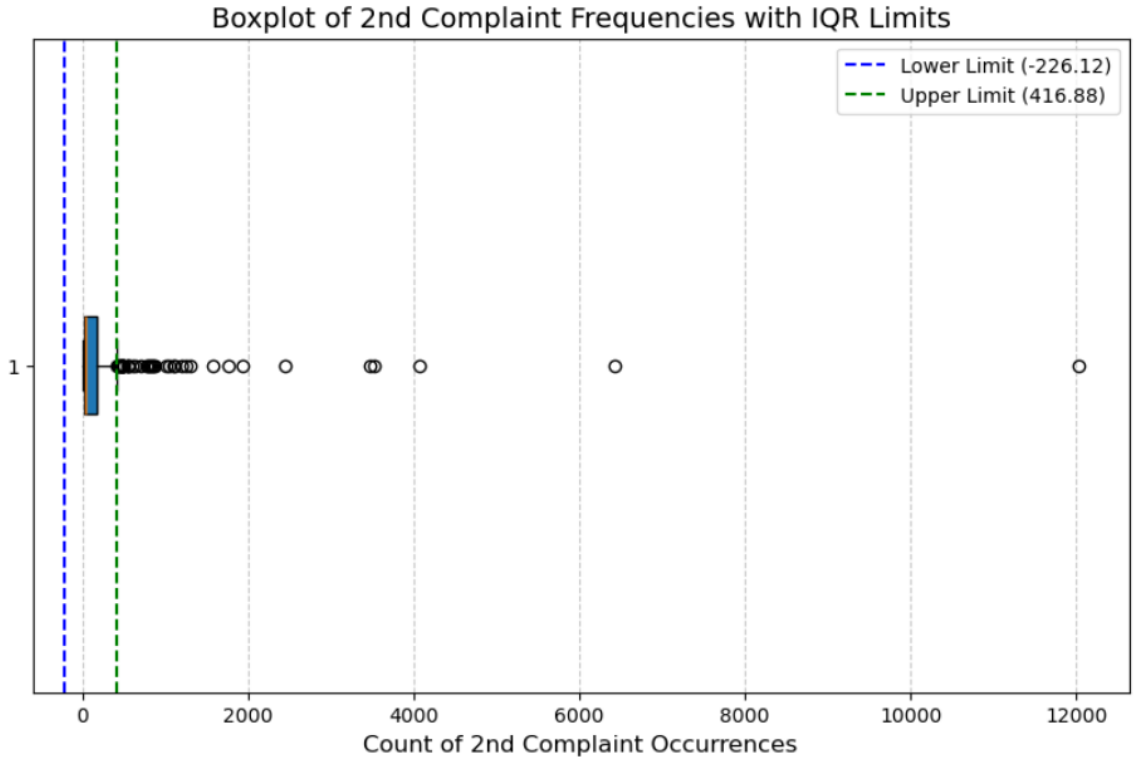

Figure 8: Boxplots of 2nd complaint frequencies showing outliers.

	3rd Complaint	Count
0	fever	8855
1	chills	6994
2	headache	5863
3	bodyaches	4285
4	nausea	4194
...	...	...
246	blood vomit	2
247	Fb throat	2
248	fracture	2
249	pain ear	1
250	Anorexia	1

251 rows × 2 columns

Figure 9: Count of 3rd Complaint

```
Q1: 11.0
Q2 (Median): 28.0
Q3: 143.5
IQR: 132.5
Lower Limit: -187.75
Upper Limit: 342.25
Number of outliers: 251
```

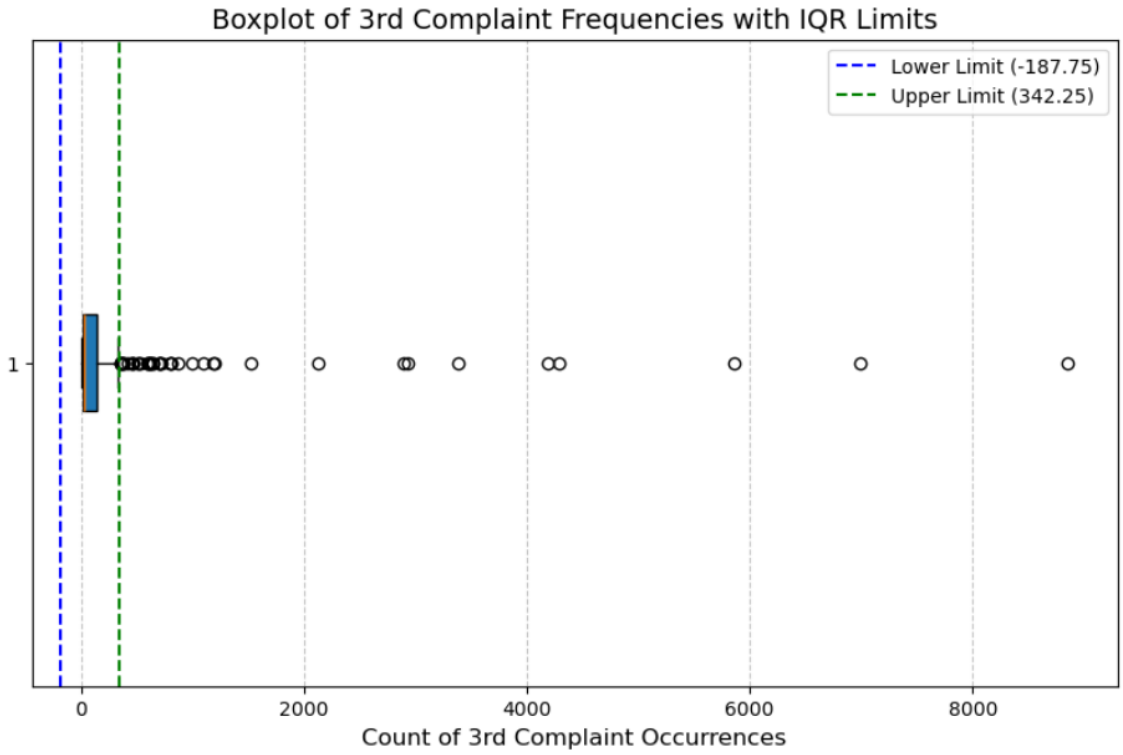

Figure 10: Boxplots of 3rd complaint frequencies showing outliers.

```
Q1: 22.0
Q2 (Median): 30.0
Q3: 45.0
IQR: 23.0
Lower Limit: -12.5
Upper Limit: 79.5
Number of outliers: 32
```

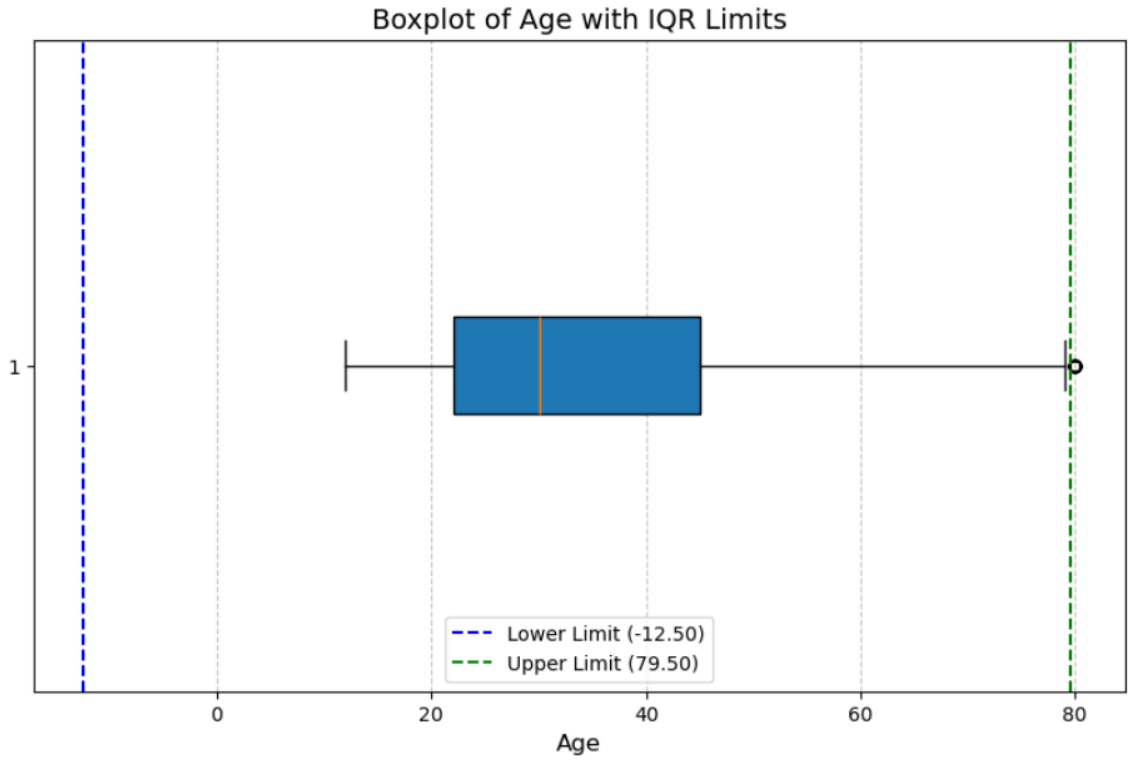

Figure 11: Boxplot of age showing outliers.

6.1.2 Parameters

- **Period:** Defines the cycle (e.g., 7 for weekly periodicity).
- **Seasonal and Trend Smoothing:** Control the smoothness of the decomposition.

6.2 Stationarity and Augmented Dickey-Fuller (ADF) Test

Stationarity ensures that a time series has consistent statistical properties (mean, variance) over time. ADF test checks for stationarity:

$$\Delta y_t = \beta y_{t-1} + \alpha t + \gamma \Delta y_{t-1} + \epsilon_t$$

- H_0: The time series is non-stationary.
- H_1: The time series is stationary.

If the p-value < 0.05, H_0 is rejected, indicating stationarity.

6.3 Anomaly Detection Using STL

Using the residual component (R_t):

- Compute $Q1$, $Q3$, and IQR (IQR $= Q3 - Q1$).

Converting Date column into Date time format

```
df['Date'] = pd.to_datetime(df['Date'], dayfirst=True, errors='coerce')
```

```
df
```

	Date	Age	Sex	1st complaint	2nd Complaint	3rd Complaint	Address	Latitude	Longi
0	2011-05-01	18	Male	fever	bodyaches	Headache	KOT ABDUL MALIK	31.620420	74.23
1	2011-05-01	20	Male	pain	bleeding	RTA	Gulshan-e-Ravi	31.552170	74.27
2	2011-05-01	40	Male	Dyspnea	chest pain	RTI	Gawalmandi	31.571870	74.31
3	2011-05-01	24	Male	headache	allergy	Vomiting	KOT ABDUL MALIK	31.620420	74.23
6	2011-05-01	20	Male	pain	bleeding	RTA	Gawalmandi	31.571870	74.31
...	...	...	...	...	...	...	...	...	
83452	2011-08-31	31	Female	Fever	Chills	Body pains	Gujranwala	32.166351	74.19
83453	2011-08-31	35	Female	Vomiting	fever	nausea	Anarkali Bazaar Lahore	31.569800	74.31
83454	2011-08-31	35	Female	Arthritis	Anxiety	arthralgia	Gawalmandi	31.571870	74.31
83456	2011-08-31	46	Male	Palpitation	anxiety	arthralgia	Lahori Gate	31.577410	74.31
83457	2011-08-31	18	Female	Fever	headache	chills	Gawalmandi	31.571870	74.31

74293 rows × 10 columns

Figure 12: Conversion of *Date* column to datetime format.

- Define thresholds:

$$\text{Lower Limit} = Q1 - 1.5 \times \text{IQR}$$
$$\text{Upper Limit} = Q3 + 1.5 \times \text{IQR}$$

- Identify anomalies as residuals outside these limits.

See Figures 26-31.

—

7 Data Aggregation Using K-Means Clustering

7.1 Mathematics of K-Means

K-Means minimizes the within-cluster variance:

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Where:

- k: Number of clusters.
- C_i: Cluster i.
- μ_i: Centroid of cluster C_i.

	Date	Age	Sex	1st complaint	2nd Complaint	3rd Complaint	Address	Latitude	Longitude	Age_Group
0	01.05.2011	18	Male	fever	bodyaches	Headache	KOT ABDUL MALIK	31.620420	74.234381	[10, 20)
1	01.05.2011	20	Male	pain	bleeding	RTA	Gulshan-e-Ravi	31.552170	74.275290	[20, 30)
2	01.05.2011	40	Male	Dyspnea	chest pain	RTI	Gawalmandi	31.571870	74.318260	[40, 50)
3	01.05.2011	24	Male	headache	allergy	Vomiting	KOT ABDUL MALIK	31.620420	74.234381	[20, 30)
6	01.05.2011	20	Male	pain	bleeding	RTA	Gawalmandi	31.571870	74.318260	[20, 30)
...	...	...	...	...	...	...	...	...	...	...
83452	31.08.2011	31	Female	Fever	Chills	Body pains	Gujranwala	32.166351	74.195900	[30, 40)
83453	31.08.2011	35	Female	Vomiting	fever	nausea	Anarkali Bazaar Lahore	31.569800	74.312000	[30, 40)
83454	31.08.2011	35	Female	Arthritis	Anxiety	arthralgia	Gawalmandi	31.571870	74.318260	[30, 40)
83456	31.08.2011	46	Male	Palpitation	anxiety	arthralgia	Lahori Gate	31.577410	74.313430	[40, 50)
83457	31.08.2011	18	Female	Fever	headache	chills	Gawalmandi	31.571870	74.318260	[10, 20)

'4293 rows × 10 columns

Figure 13: Creation of new features such as *Latitude*, *Longitude*, and *Age Group*.

7.2 Elbow Method

The elbow method identifies the optimal number of clusters (k) by plotting:

$$\text{Inertia} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

The "elbow point" is where inertia starts to decrease slowly.
See Figure 32.

7.3 Data Aggregation for Gender and Month

- Subset data by gender (male, female) and month (May, June, July, August).
- Perform clustering on Latitude and Longitude for each subset.
- Visualize clusters using Folium maps.

See Figures 33-40.

8 Predictive Analysis

8.1 Problem Statement

The objective of this analysis is to develop a machine learning model to predict:

- Total number of males and females registered on each date.
- Age group distributions for males and females.

The predictive model aims to assist in identifying trends in patient demographics and preparing resources effectively.

	Latitude	Longitude	total_patients	normalized_patients
0	30.813802	73.453378	195	0.002395
1	31.449151	73.712479	104	0.000000
2	31.460495	74.221700	114	0.000263
3	31.531300	74.318300	309	0.005396
4	31.536960	74.271942	719	0.016188
5	31.549700	74.343600	1811	0.044932
6	31.552170	74.275290	539	0.011450
7	31.552830	74.295688	119	0.000395
8	31.559700	74.313800	924	0.021584

Figure 14: Normalized patient counts for hotspot visualization.

	Address	total_males	total_females	major_1st_complaint	major_2nd_complaint	major_3rd_complaint	total_patients
0	Anarkali Bazaar Lahore	471	353	fever	fever	fever	824
1	Badami Bagh	1259	1042	fever	fever	fever	2301
2	Baghbanpura	209	156	fever	fever	chills	365
3	Band Rd Lahore	386	333	fever	fever	fever	719

Figure 15: Grouped data with calculated total cases based upon locations

8.2 Data Preparation

- **Aggregating Data:** The dataset was grouped by *Date*, calculating:
  - Total number of males and females for each day.
  - Age group distributions for both genders.
- **Feature Engineering:** The *Date* column was transformed into ordinal format to facilitate regression modeling.
- **Data Splitting:** The dataset was divided into **training** (80%) and **testing** (20%) subsets to evaluate model performance.

8.3 Model Selection

Two machine learning models were implemented for this regression task:

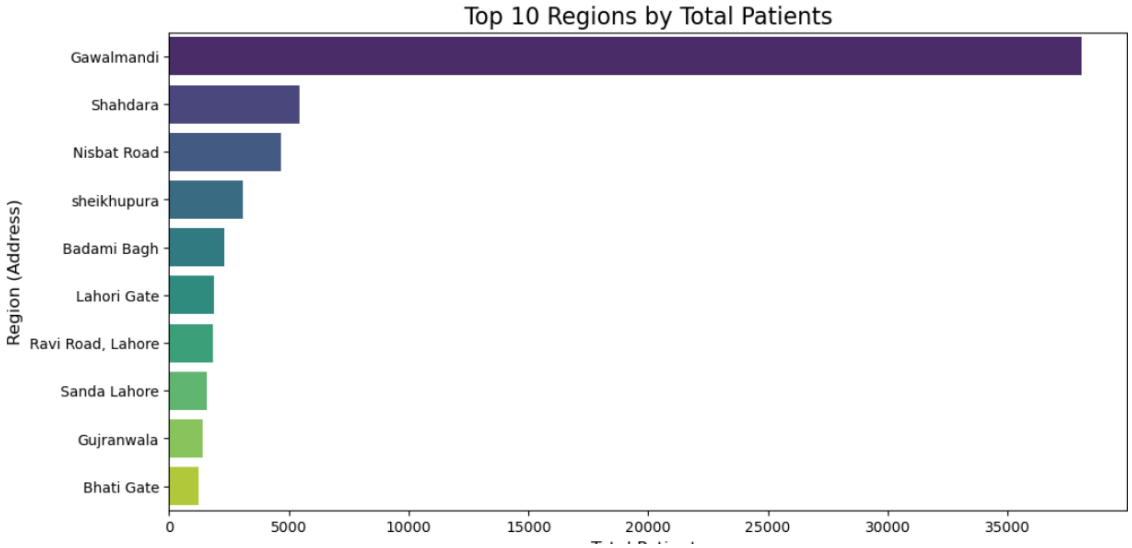

Figure 16: Top 10 Address Locations with highest number of patients

- **Random Forest Regression:** A multi-output regressor was employed to predict multiple target variables, including total males, total females, and age group distributions. This model is well-suited for capturing nonlinear relationships.
- **Support Vector Regression (SVR):** Using the Radial Basis Function (RBF) kernel, this model maps input features to higher-dimensional spaces for better regression accuracy.

8.4 Model Evaluation

The models were evaluated using the following metrics:

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors.
- **Mean Squared Error (MSE):** Penalizes larger errors more heavily.
- **R-Squared ($\mathbf{R^2}$):** Assesses how well the model explains the variance in the data.

Random Forest Results:

- MAE: `[110.41, 91.04, ...]` for different targets.
- MSE: `[33672.26, 32838.78, ...]` for different targets.
- R^2 Score: Ranging from -0.577 (for some targets) to 1.0 (perfect predictions).

Support Vector Regression Results:

- **Hyperparameter Tuning:**
  - Optimized parameters using **RandomizedSearchCV**:
    - $C = 98.77$
    - $\gamma = 0.01879$
- Model Performance: Model Performance w.r.t R^2 Score of Random Forest and SVM is given in Table 1.

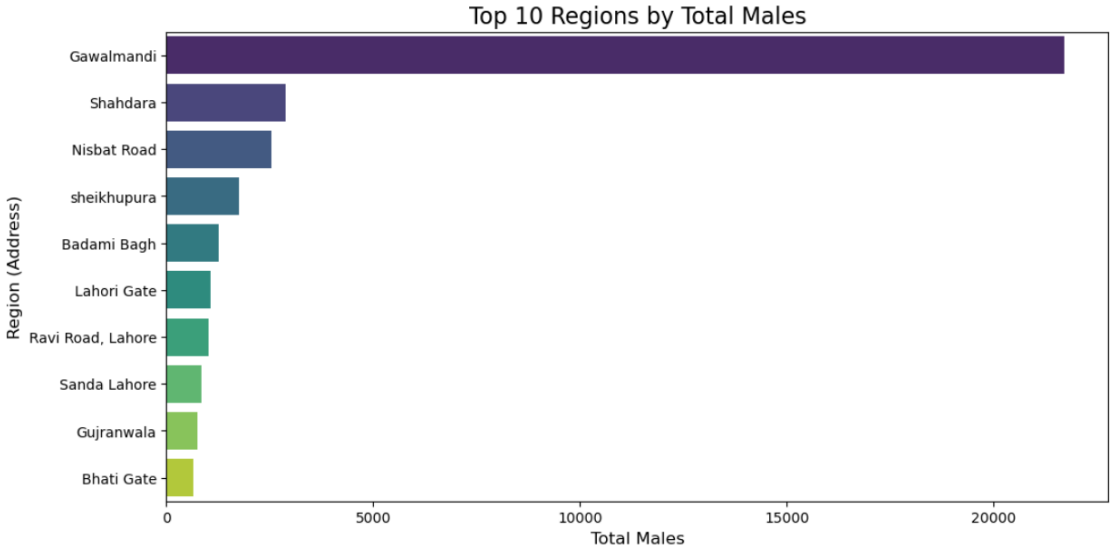

Figure 17: Grouped data with calculated total males based upon locations

Table 1: R-Squared (R^2) Scores for Random Forest and SVM Across Targets

Target	Random Forest R^2 Score	SVM R^2 Score
Total Males	0.517	0.244
Total Females	-0.530	0.465
[0, 10) Males	1.000	1.000
[10, 20) Males	0.938	0.502
[20, 30) Males	0.723	0.343
[30, 40) Males	-0.050	0.388
[40, 50) Males	0.202	0.456
[50, 60) Males	0.024	0.406
[60, 70) Males	-0.143	0.251
[70, 80) Males	-0.618	0.283
[80, 90) Males	-0.371	-0.181
[0, 10) Females	1.000	1.000
[10, 20) Females	1.000	1.000
[20, 30) Females	-1.177	0.576
[30, 40) Females	-0.749	0.522
[40, 50) Females	0.198	0.446
[50, 60) Females	-0.293	0.248
[60, 70) Females	0.116	0.307
[70, 80) Females	0.250	0.377
[80, 90) Females	-0.322	0.225
[90, 100) Females	0.202	0.041

8.5 Model Optimization

- **Hyperparameter Tuning:** The SVR model's C (regularization parameter) and γ (kernel coefficient) were optimized using **RandomizedSearchCV**.
- **Cross-Validation:** A 3-fold cross-validation approach ensured robust evaluation of the models.

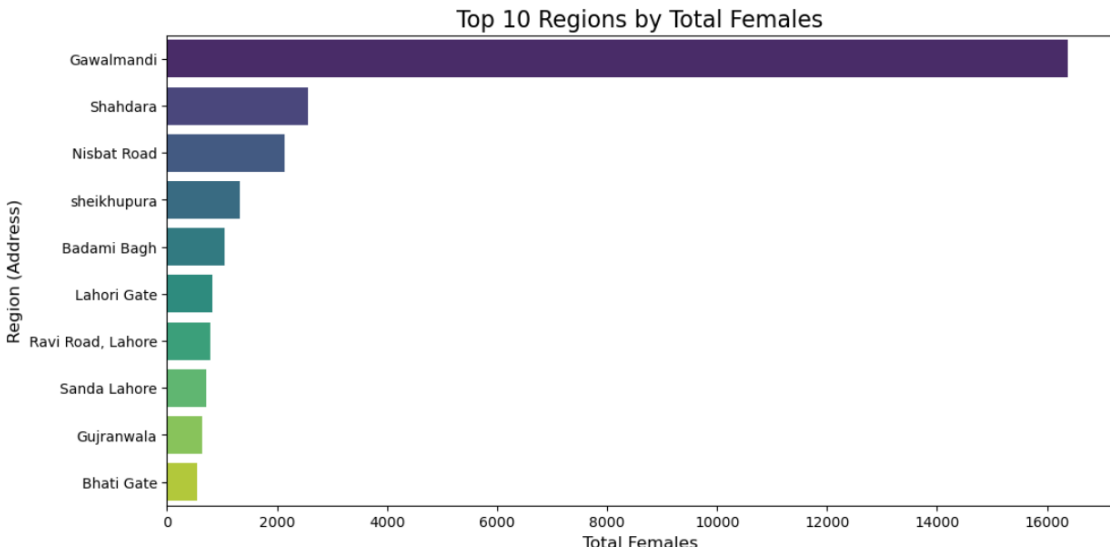

Figure 18: Grouped data with calculated total females based upon locations

9 Predictive Analysis: Log Transformation for Model Optimization

9.1 Introduction

Log transformation is an effective method to optimize model performance when working with datasets containing outliers or highly skewed distributions. It helps stabilize variance, reduces the effect of extreme values, and improves the model's ability to generalize better to unseen data.

9.2 Log Transformation Application

Before training the models, a **log transformation** was applied to all target variables, including the `total_males`, `total_females`, and their corresponding age groups. This transformation was used to reduce the effect of outliers and make the data more suitable for machine learning models.

Why Use Log Transformation?

- Reduces skewness by compressing large values and spreading out small ones.
- Mitigates the impact of outliers by compressing extreme values.
- Linearizes non-linear relationships, making it easier for models to identify patterns.
- Helps meet the assumptions of linear models, which often perform better with normally distributed data.

The following code demonstrates the application of log transformation to all target variables:

```
import numpy as np

# Apply log transformation to all target columns
for col in target_columns:
    final_data[col] = np.log1p(final_data[col])  # log1p handles zeros
```

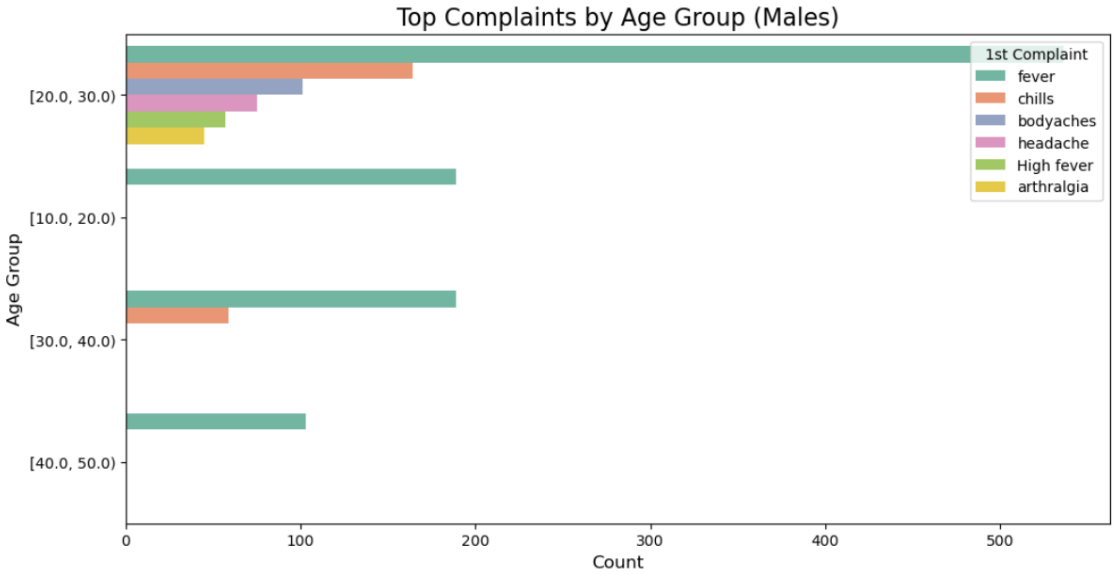

Figure 19: Displays the top 10 complaint trends with a grouped bar chart (Males, 1st Complaint)

9.3 Impact on Model Performance

The application of log transformation resulted in more stable and accurate predictions. After log transformation, the model was better equipped to handle skewed distributions and outliers, leading to improvements in the predictive accuracy.

10 Results and Discussion

10.1 R-squared (R^2) Comparison for Random Forest vs. SVM

The following table compares the R-squared (R^2) values for both the Random Forest and SVM models for each target variable.

10.2 Discussion

10.2.1 Random Forest Performance:

The **Random Forest** model showed generally strong performance across all target variables. It achieved high R^2 values, especially for `total_males` (0.861) and `total_females` (0.827). This indicates that the Random Forest model effectively captured the relationships between the date and the number of males and females in each age group. The highest R^2 of 1.000 for the **[0, 10)** age group suggests perfect predictions for this group.

However, some age groups, such as **[70, 80)** and **[80, 90)** for both males and females, showed lower R^2 values, suggesting that the model was less able to predict these groups. This could be due to a smaller sample size or a lack of clear patterns for these groups.

10.2.2 SVM Performance:

The **SVM** model also performed well, though it showed slightly lower R^2 values for most targets compared to Random Forest. For `total_males` and `total_females`, the R^2 values were 0.837 and 0.722, respectively, which are lower than those of Random Forest.

SVM performed well in some age groups, particularly **[0, 10)** and **[10, 20)**, where it achieved perfect R^2 values of 1.000. However, it showed weaker performance in other age groups,

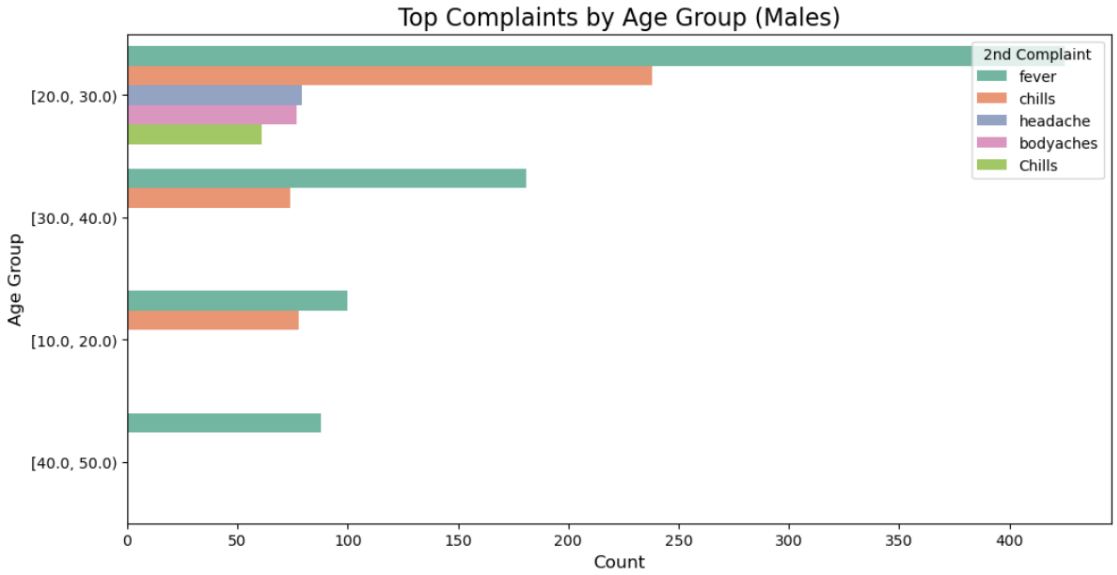

Figure 20: Displays the top 10 complaint trends with a grouped bar chart (Males, 2nd Complaint)

Target	Random Forest R^2	SVM R^2
Total Males	0.861	0.837
Total Females	0.827	0.722
[0, 10) (Age Group Males)	1.000	1.000
[10, 20) (Age Group Males)	0.984	0.840
[20, 30) (Age Group Males)	0.809	0.675
[30, 40) (Age Group Males)	0.723	0.748
[40, 50) (Age Group Males)	0.723	0.763
[50, 60) (Age Group Males)	0.613	0.592
[60, 70) (Age Group Males)	0.458	0.505
[70, 80) (Age Group Males)	0.324	0.519
[0, 10) (Age Group Females)	-0.371	-0.117
[10, 20) (Age Group Females)	1.000	1.000
[20, 30) (Age Group Females)	1.000	1.000
[30, 40) (Age Group Females)	0.755	0.721
[40, 50) (Age Group Females)	0.706	0.712
[50, 60) (Age Group Females)	0.669	0.677
[60, 70) (Age Group Females)	0.674	0.610
[70, 80) (Age Group Females)	0.448	0.556
[80, 90) (Age Group Females)	0.450	0.550
[90, 100) (Age Group Females)	0.209	0.468

Table 2: R^2 for Each Target - Random Forest vs. SVM

such as **[0, 10)** (Females) and **[70, 80)** (Females), where negative or low R^2 values were observed. These discrepancies may result from SVM's sensitivity to data distributions, making it less robust in some cases compared to Random Forest.

10.3 Conclusion

Both the **Random Forest** and **SVM** models performed reasonably well for predicting the total counts of males and females, as well as the age group distributions. However, **Random

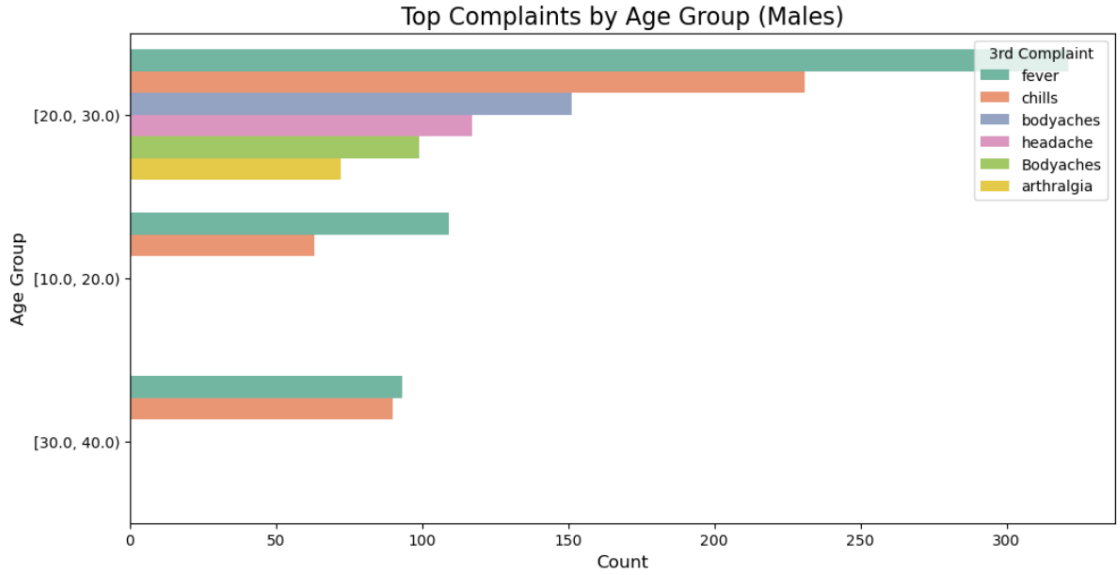

Figure 21: Displays the top 10 complaint trends with a grouped bar chart (Males, 3rd Complaint)

Forest** outperformed **SVM** in terms of R^2 values for most targets. The higher R^2 values in Random Forest indicate that it better captures the relationships between date and the targets (total males, total females, and age group distributions).

While **SVM** performed well for some specific age groups, it showed lower predictive power in others, especially for smaller groups. MSE and MAE for Random Forest for targets, Total males and males are lower than SVR (Figures 41 and 42). Therefore, for this particular problem, **Random Forest** is the more reliable model for predicting the total number of males and females and their corresponding age groups.

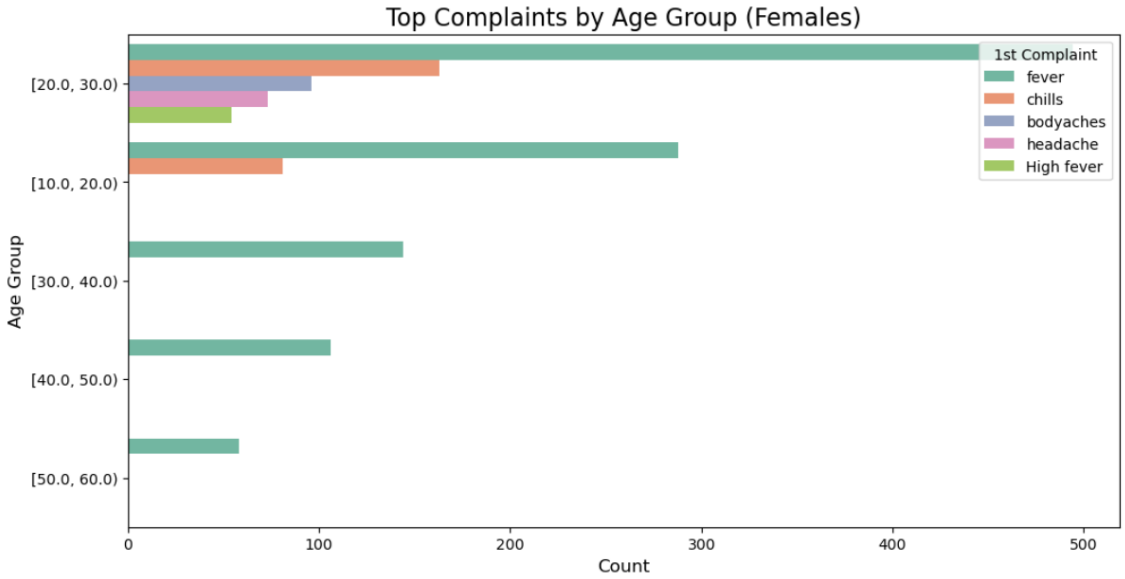

Figure 22: Displays the top 10 complaint trends with a grouped bar chart (Females, 1st Complaint)

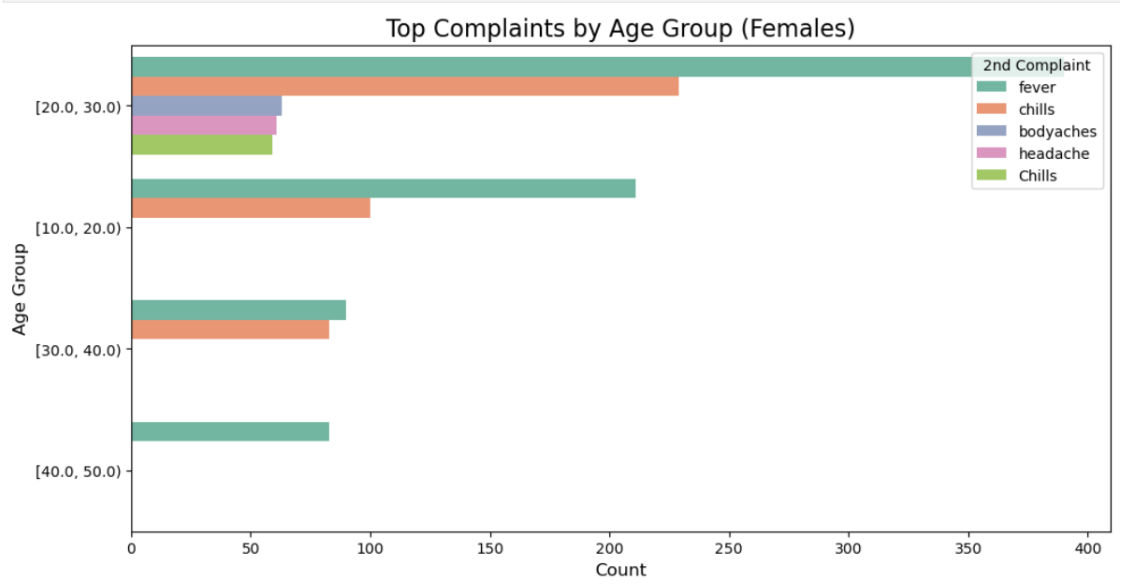

Figure 23: Displays the top 10 complaint trends with a grouped bar chart (Females, 2nd Complaint)

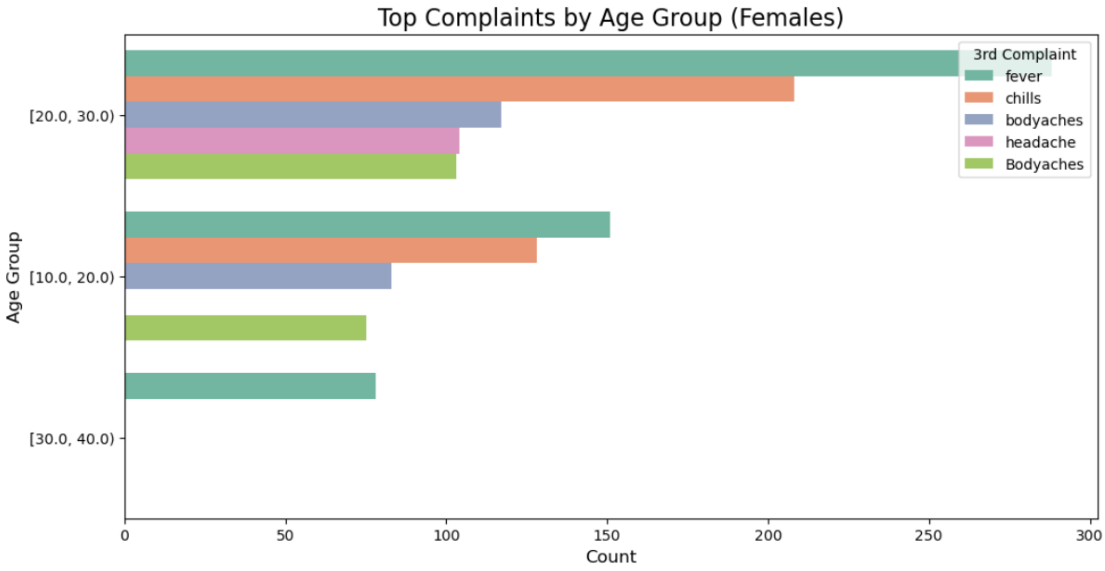

Figure 24: Displays the top 10 complaint trends with a grouped bar chart (Females, 3rd Complaint)

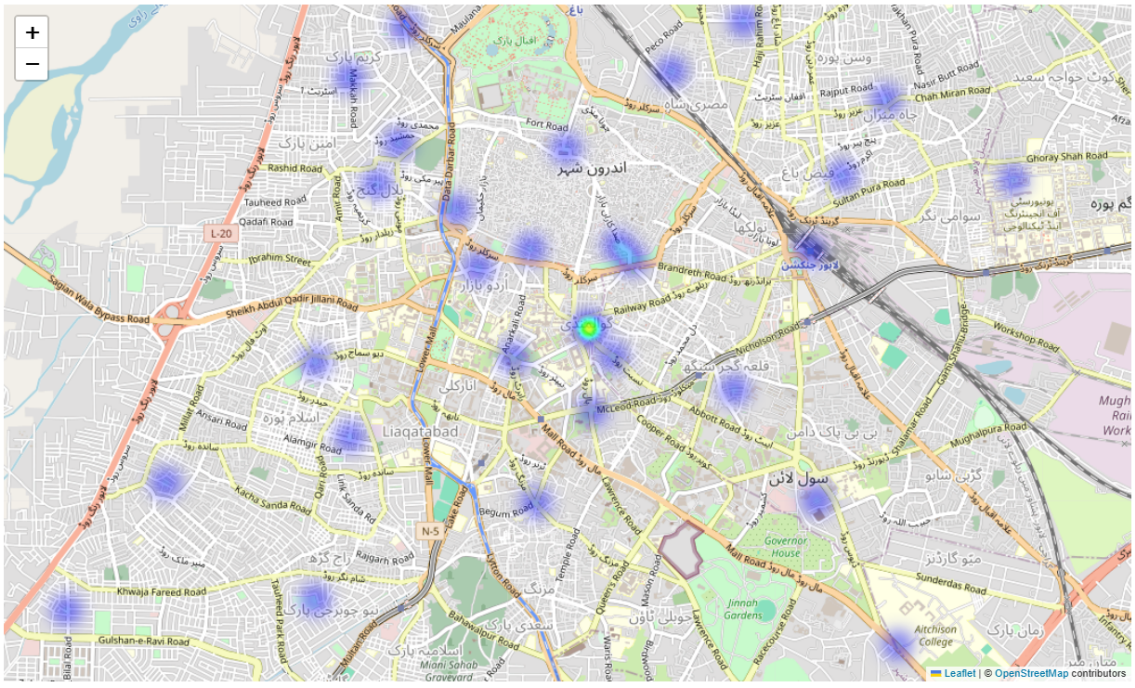

Figure 25: Hotspots

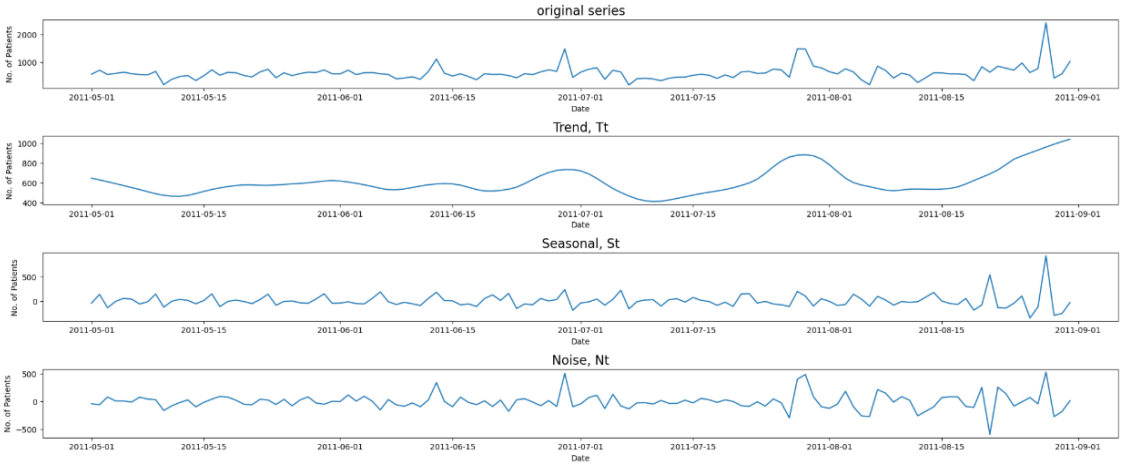

Figure 26: Time Series Total Patients

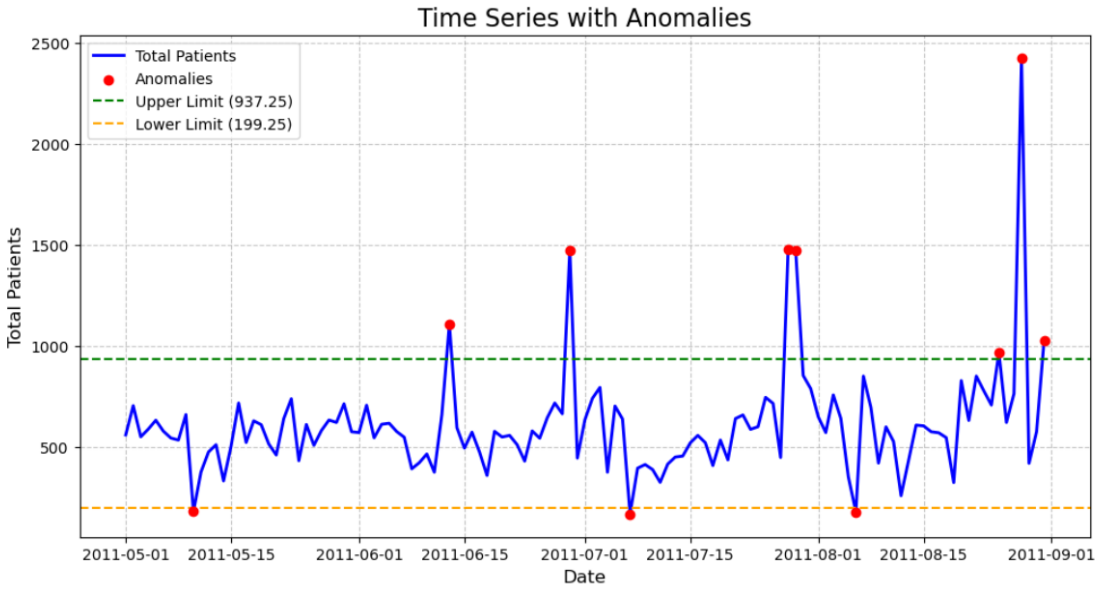

Figure 27: Time Series Total Patients

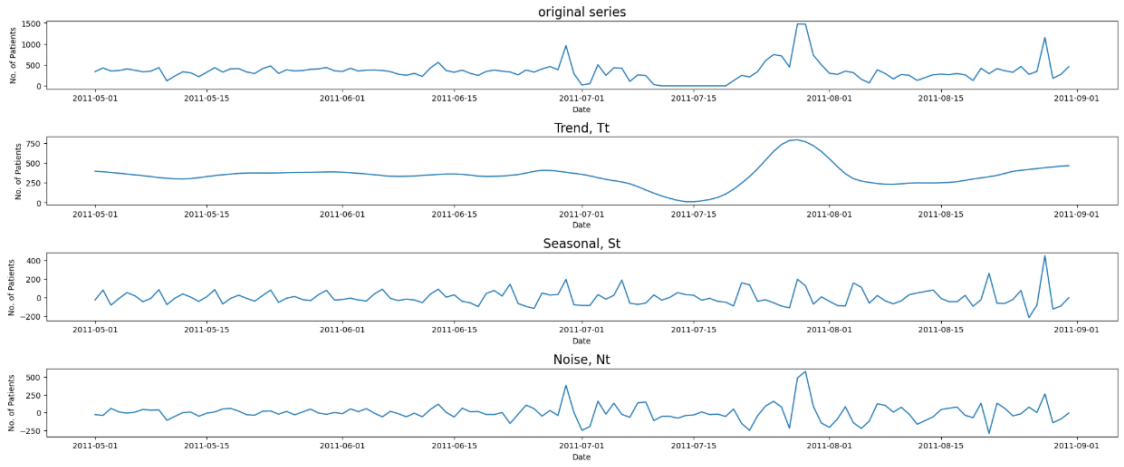

Figure 28: Time Series Total males

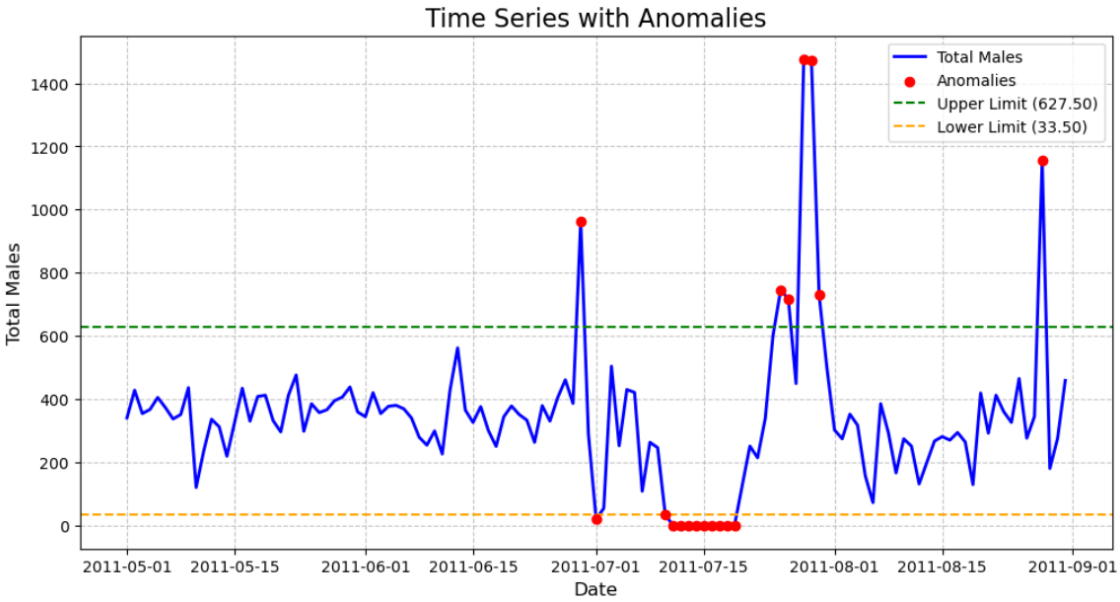

Figure 29: Time Series Total males

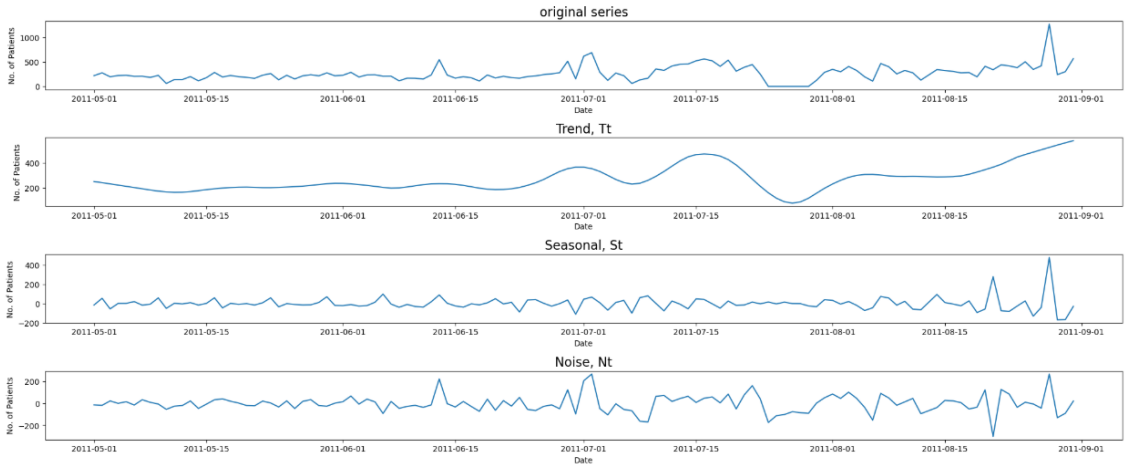

Figure 30: Time Series Total females

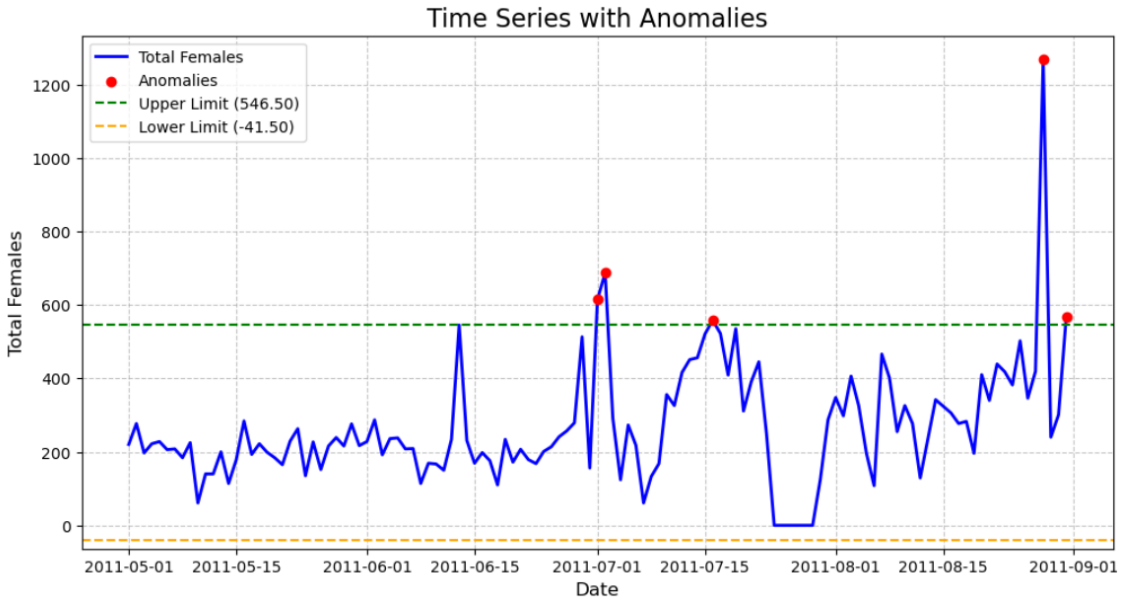

Figure 31: Time Series Total females

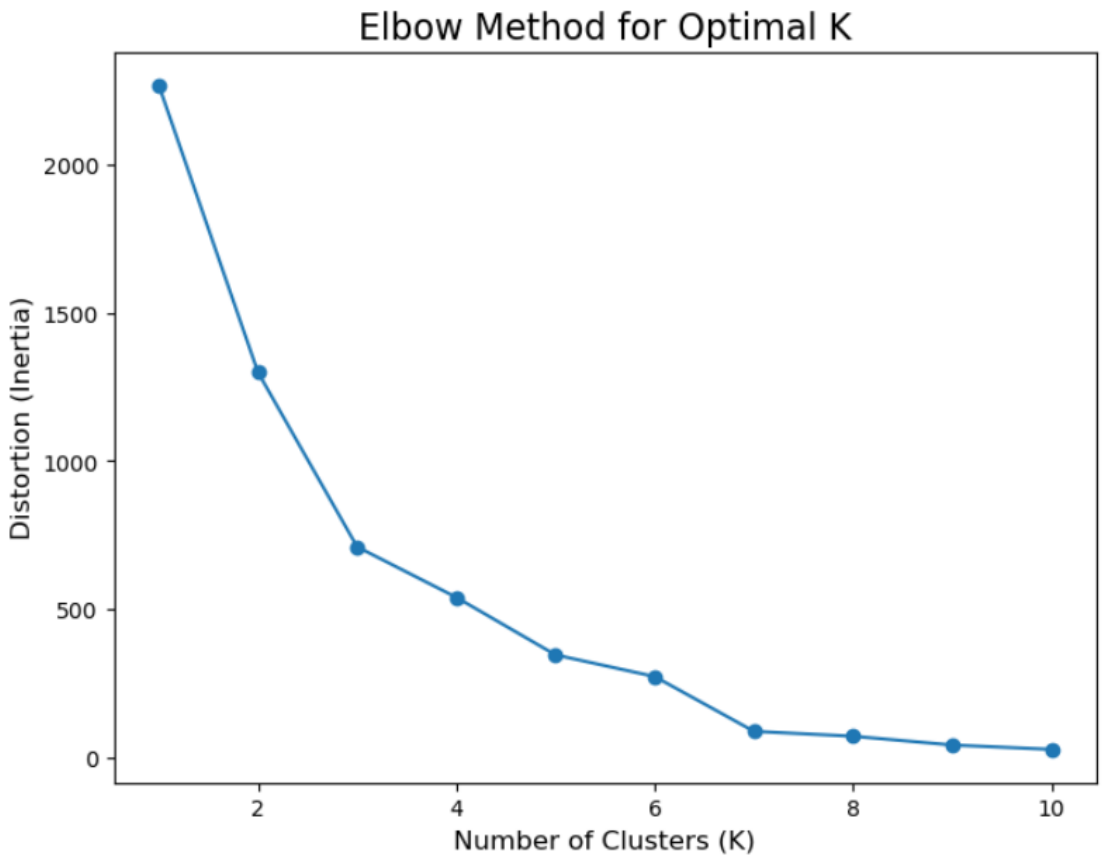

Figure 32: Elbow Method

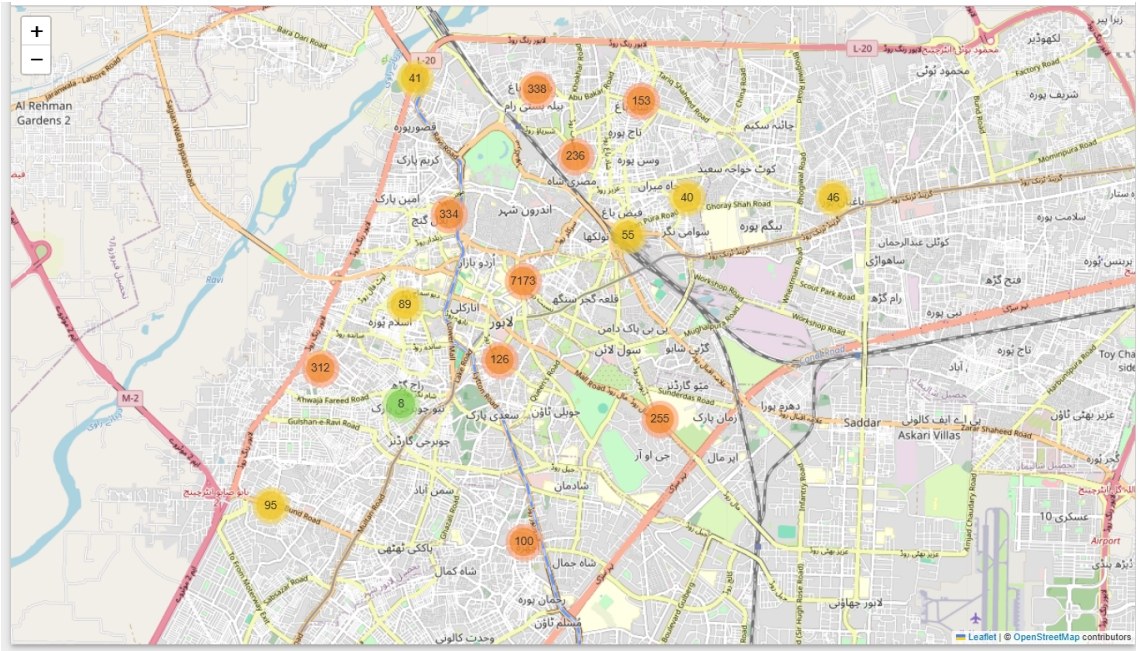

Figure 33: Males-May

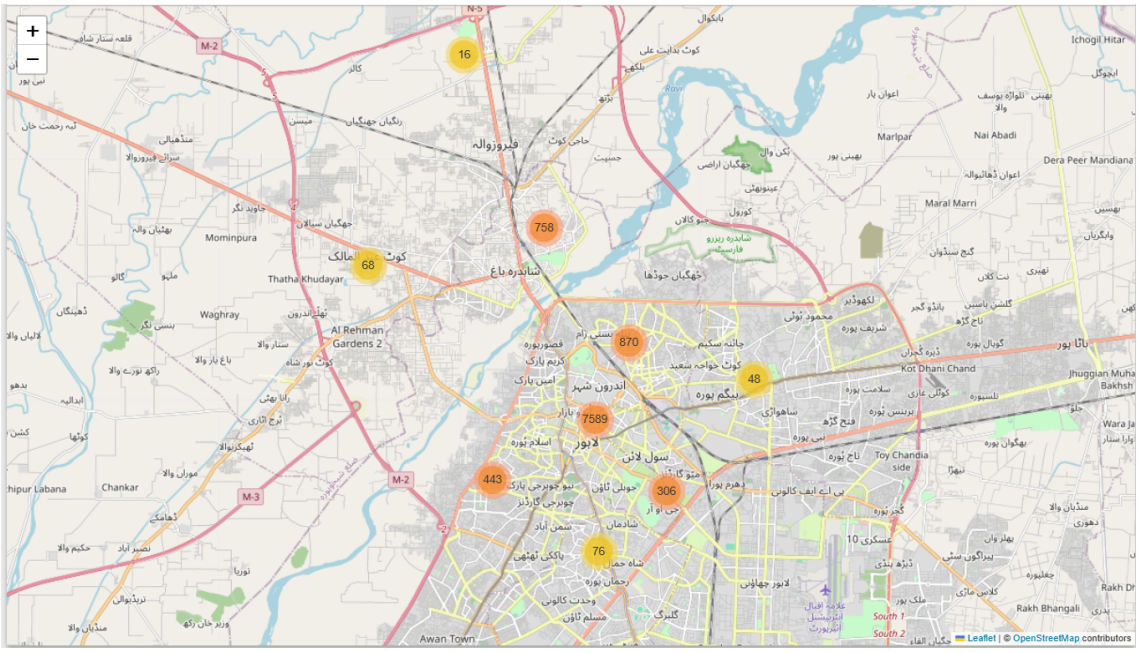

Figure 34: Males-June

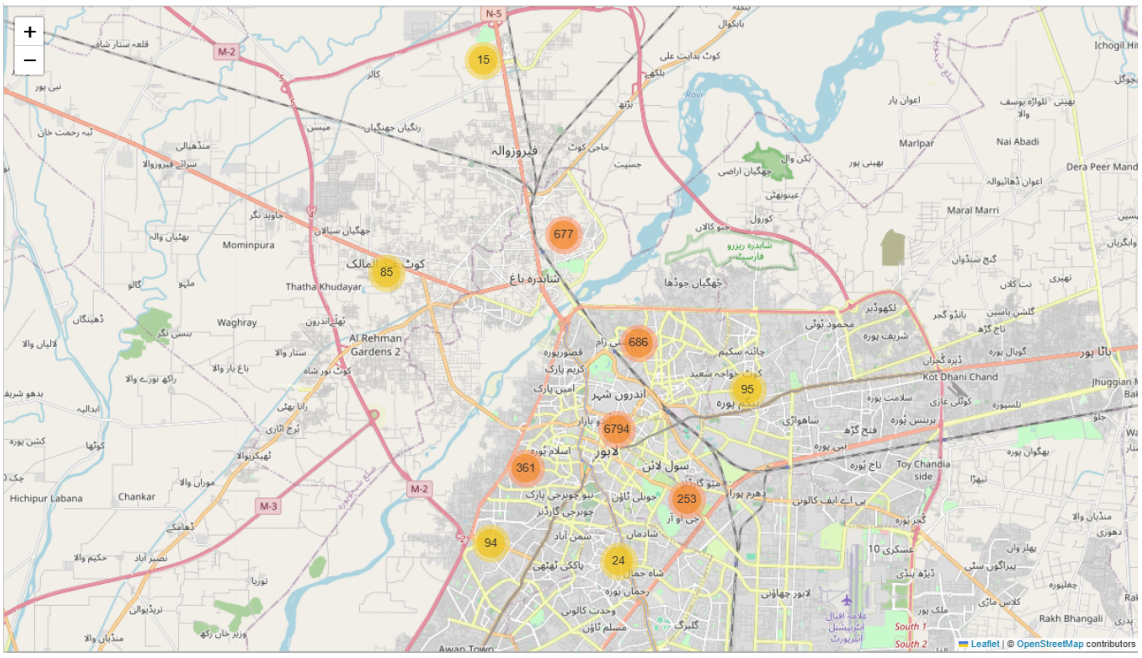

Figure 35: Males-July

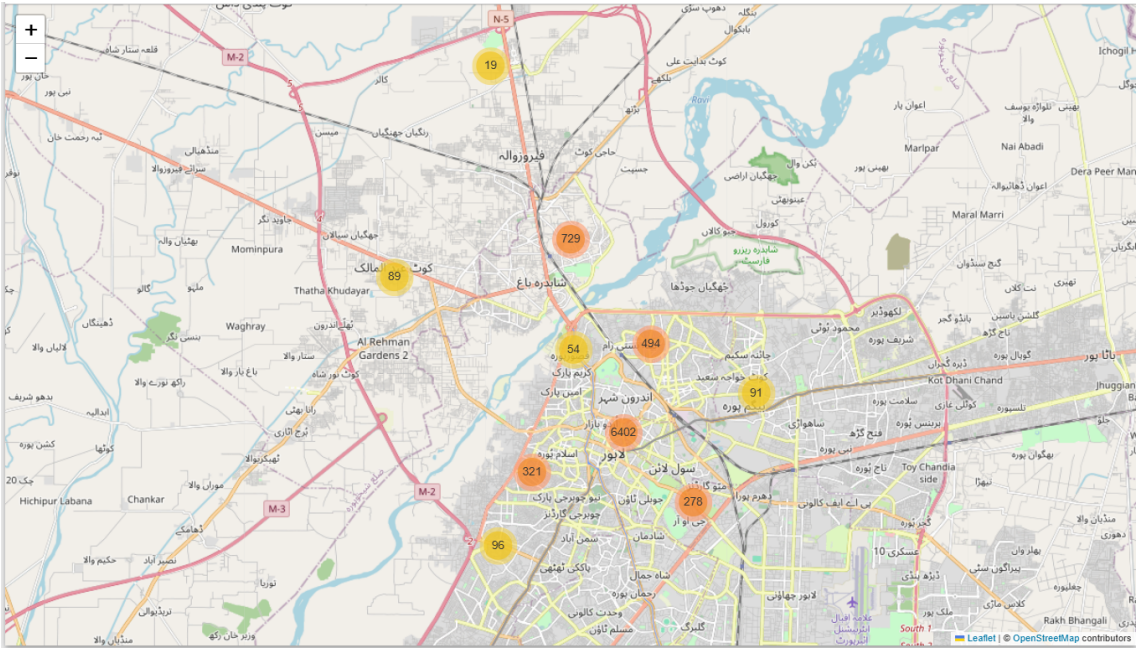

Figure 36: Males-Aug

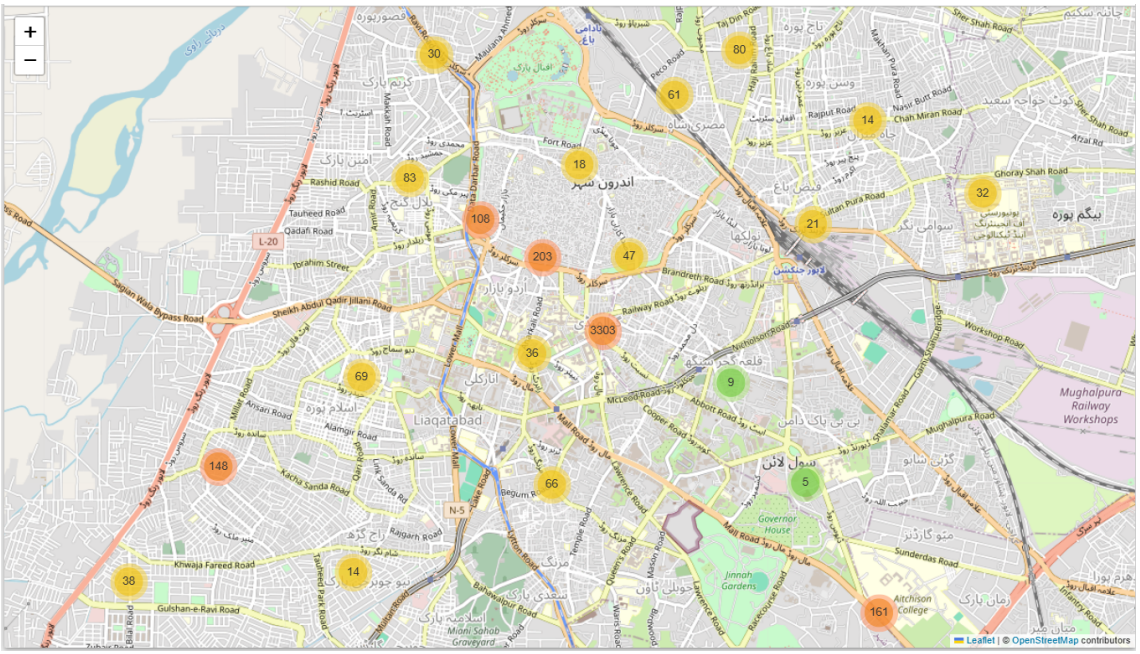

Figure 37: Females-May

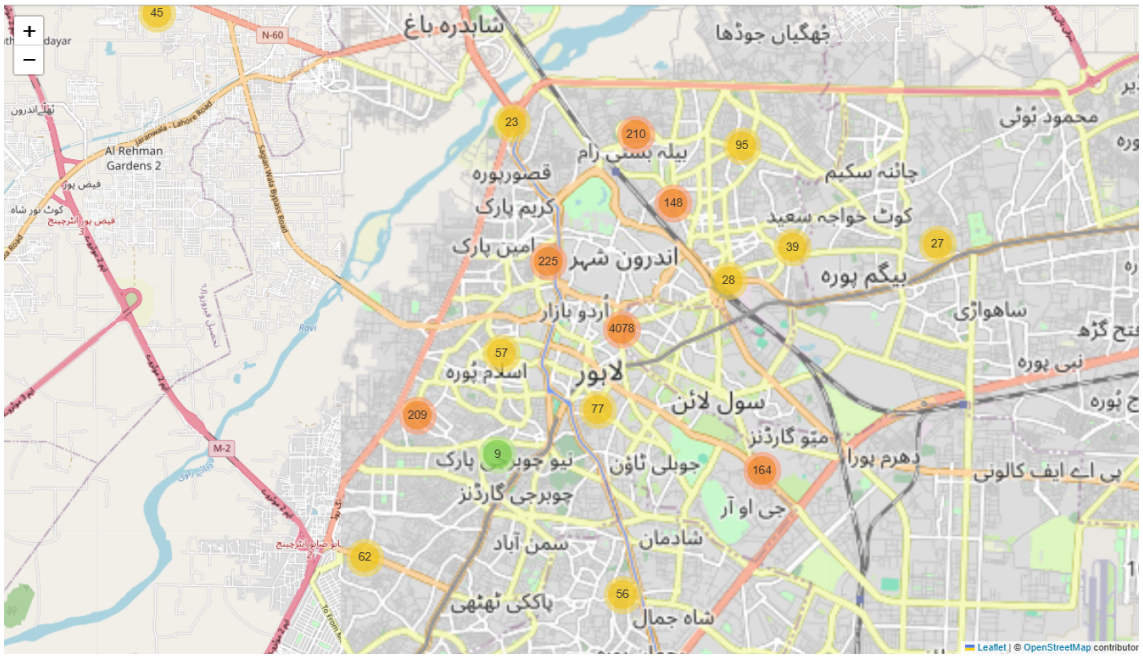

Figure 38: Females-June

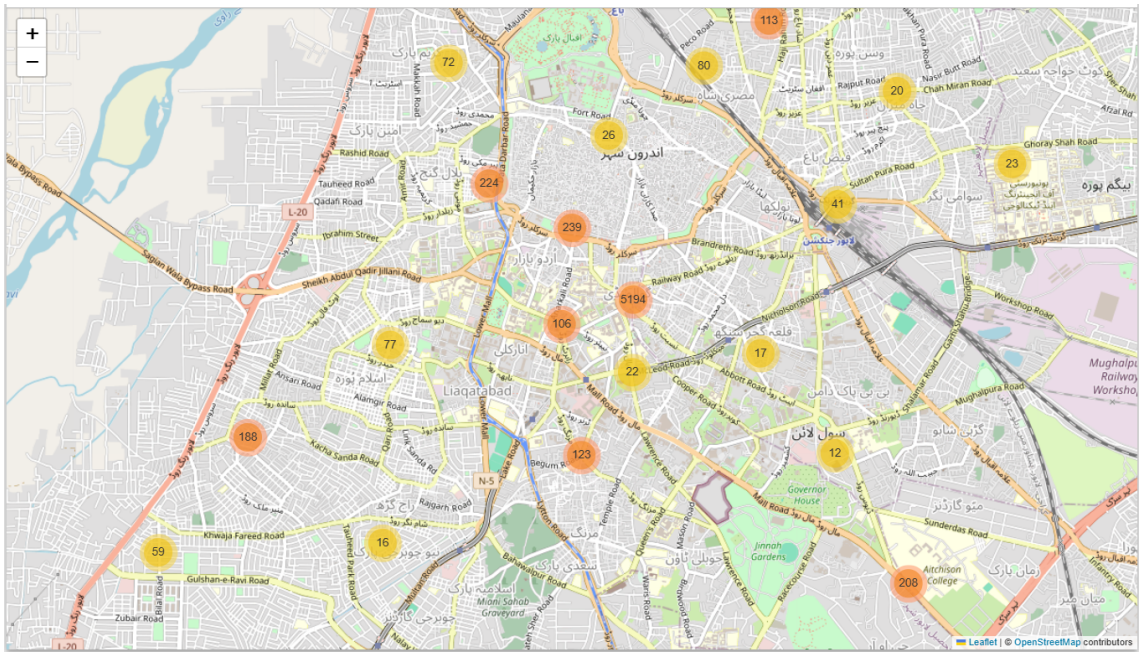

Figure 39: Females-July

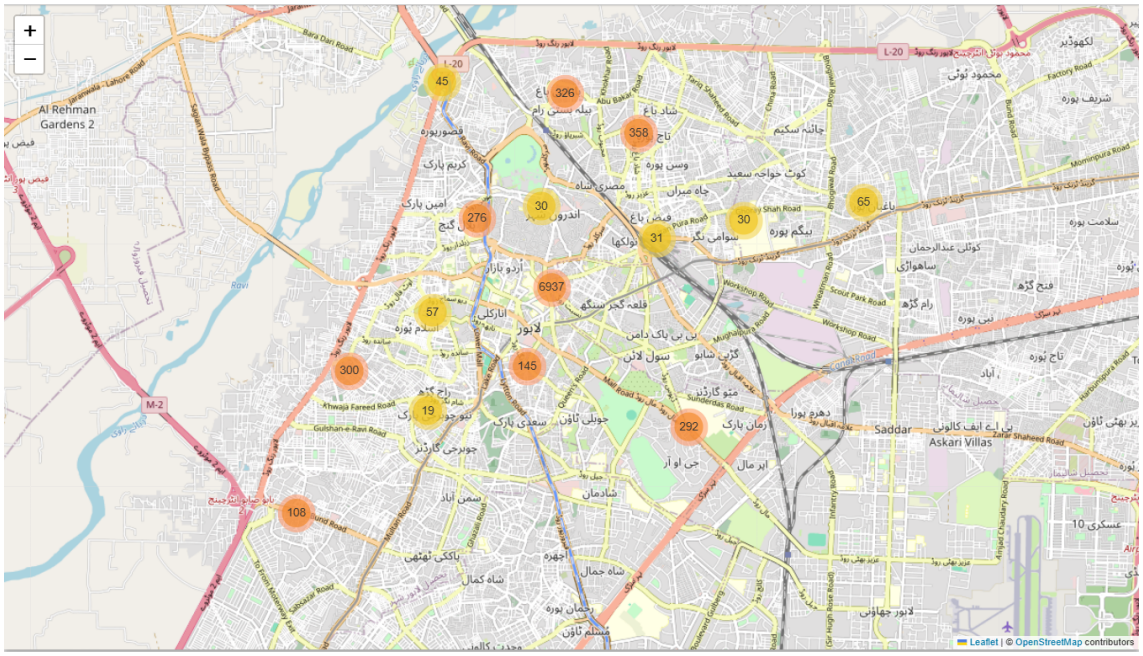

Figure 40: Females-Aug

Target	R2 (Random Forest)	R2 (SVR)	MAE (Random Forest)	MAE (SVR)	MSE (Random Forest)	MSE (SVR)
Total Males	0.861	0.837	0.306	0.35	0.222	0.261
Total Females	0.827	0.722	0.323	0.394	0.228	0.366
[0, 10) (Age Group Males)	1	1	0	0	0	0
[10, 20) (Age Group Males)	0.984	0.84	0.154	0.354	0.049	0.501
[20, 30) (Age Group Males)	0.809	0.675	0.345	0.432	0.239	0.408
[30, 40) (Age Group Males)	0.723	0.748	0.331	0.327	0.265	0.241
[40, 50) (Age Group Males)	0.723	0.763	0.342	0.331	0.233	0.2
[50, 60) (Age Group Males)	0.613	0.592	0.4	0.406	0.29	0.306
[60, 70) (Age Group Males)	0.458	0.505	0.431	0.427	0.387	0.353
[70, 80) (Age Group Males)	0.324	0.519	0.451	0.419	0.432	0.307
[0, 10) (Age Group Females)	-0.371	-0.117	0.213	0.207	0.294	0.239

Figure 41: R2, MSE, MAE Comparison for Random Forest and SVR

[10, 20) (Age Group Females)	1	1	0	0	0	0
[20, 30) (Age Group Females)	1	1	0	0	0	0
[30, 40) (Age Group Females)	0.755	0.721	0.335	0.358	0.209	0.239
[40, 50) (Age Group Females)	0.706	0.712	0.389	0.359	0.263	0.258
[50, 60) (Age Group Females)	0.669	0.677	0.391	0.379	0.234	0.229
[60, 70) (Age Group Females)	0.674	0.61	0.367	0.397	0.336	0.259
[70, 80) (Age Group Females)	0.448	0.556	0.433	0.414	0.258	0.27
[80, 90) (Age Group Females)	0.45	0.55	0.357	0.368	0.254	0.211
[90, 100) (Age Group Females)	0.209	0.468	0.044	0.099	0.022	0.031

Figure 42: R2, MSE, MAE Comparison for Random Forest and SVR