Pest Data Project Analysis Report

Introduction

This report explores a dataset on pest populations collected across various locations in India from 1959 to 2011. The data includes observations for several pest species alongside weather parameters and pest counts.

□ The specific pest species include:

- Gall Midge
- Brownplanthopper
- Greenleafhopper
- LeafFolder
- Yellowstemborer
- Caseworm
- Mirid Bug
- ZigZagleafhopper
- LeafBlast
- NeckBlast

The primary objective of this analysis is to investigate potential relationships between weather conditions and pest activity for various pest species. By analyzing these factors, we hope to gain insights that could be valuable for pest management strategies in different regions of India.

Data Analysis

We imported the data into a relational database management system (RDBMS) and created a table named "PESTDATA" to store the observations.

The table schema includes columns for:

- OBSERVATION YEAR (INT): Year of observation
- STANDARD WEEK (INT): Standard week within the year (e.g., week 1, week 2)
- PESTVALUE (INT): Number of pests found per unit (e.g., per hill)
- COLLECTIONTYPE (VARCHAR(20)): Type of collection method used
- Weather parameters:
  - MAXT (DECIMAL(5,2)): Maximum Temperature in degrees Celsius
  - MINT (DECIMAL(5,2)): Minimum Temperature in degrees Celsius
  - RH1 (DECIMAL(5,2)): Relative Humidity 1 as a percentage
  - RH2 (DECIMAL(5,2)): Relative Humidity 2 as a percentage

  - RF (DECIMAL(5,2)): Rainfall in millimeters
  - WS (DECIMAL(5,2)): Wind Speed in kilometers per hour
  - SSH (DECIMAL(5,2)): Sunshine Hours
  - EVP (DECIMAL(5,2)): Evaporation in millimeters
- PESTNAME (VARCHAR(20)): Name of the pest species (e.g., Brownplanthopper)

Data Overview:

- Overview of Data

```
'OBSERVATIONYEAR','int','NO','',NULL,''

'STANDARDWEEK','int','NO','',NULL,''

'PESTVALUE','int','NO','',NULL,''

'COLLECTIONTYPE','varchar(20)','NO','',NULL,''

'MAXT','decimal(5,2)','NO','',NULL,''

'MINT','decimal(5,2)','NO','',NULL,''

'RH1','decimal(5,2)','NO','',NULL,''

'RH2','decimal(5,2)','NO','',NULL,''

'RF','decimal(5,2)','NO','',NULL,''

'WS','decimal(5,2)','NO','',NULL,''

'SSH','decimal(5,2)','NO','',NULL,''

'EVP','decimal(5,2)','NO','',NULL,''

'PESTNAME','varchar(20)','NO','',NULL,''

'LOCATION','varchar(20)','NO','',NULL,''
```

Data Cleaning :

The data cleaning process involved identifying and addressing inconsistencies or redundancies within the dataset to ensure its quality for further analysis.

There is no missing data as the data is cleaned before uploading into the SQL.

Feature Engineering for Enhanced Pest Data Analysis

The raw pest data underwent a series of transformations to improve its usability for exploring relationships between weather variables and pest activity. Here's a detailed breakdown of the feature engineering steps implemented

- **State Standardization:** The data contained city names representing locations across India. To facilitate analysis based on broader regions, a new column named "STATE" was added. This column was populated using update statements that mapped city names to their corresponding states. For example, "Cuttack" was converted to "ODISHA," ensuring consistency and enabling regional comparisons.

  ```
  'STATE','varchar(20)','NO','',NULL,''
  ```

Cuttack	**becomes**	**ODISHA**
Raipur	**becomes**	**CHHATTISGARH**
Palampur	**becomes**	**Himachal Pradesh**
Maruteru	**becomes**	**ANDHRA PRADESH**
Ludhiana	**becomes**	**PUNJAB**
Rajendranagar	**becomes**	**MANIPUR**

- **Seasonal Categorization:** A crucial factor influencing pest activity is seasonality. To capture this aspect, a new column named "SEASON" was created. This column categorized observations based on the standard week they fell within. Here's the breakdown of the seasonal mapping.

  ```
  'SEASON','varchar(20)','NO','',NULL,''
  ```

  Summer: Weeks 10 to 20 (approximate)

Monsoon: Weeks 21 to 36 (approximate)

Post-Monsoon: Weeks 37 to 44 (approximate)

Winter: Weeks 45 to 9 (wrapping around to the next year)

With this seasonal classification, we can investigate potential variations in pest populations across different seasons.

- **Interaction Term Creation:** While analyzing the independent effects of weather variables like temperature and rainfall on pest counts is valuable, exploring their combined influence can be even more insightful. To achieve this,two interaction terms were calculated and added as new columns**:**

  - **TEMP_PEST_INTERACTION (MAXT * PESTVALUE):** This term multiplies the maximum temperature (MAXT) by the pest value (PESTVALUE). This allows us to investigate if higher temperatures combined with higher initial pest populations lead to a more significant increase in pest activity compared to the individual effects of each variable.
  - **RAINFALL_PEST_INTERACTION (RF * PESTVALUE):** This term multiplies rainfall (RF) by the pest value (PESTVALUE). Similar to the temperature interaction term, this allows us to explore if higher rainfall combined with higher initial pest populations results in a more substantial**.**

    'TEMP_PEST_INTERACTION','float','NO','',NULL,''

    'RAINFALL_PEST_INTERACTION','float','NO','',NULL,''

- **Unique Record Identification:** To efficiently manage and reference individual data points, an auto-incrementing ID column named "ID" was added as the primary key for the table. This unique identifier simplifies record retrieval and manipulation within the database.

  'ID','int','NO','PRI',NULL,'auto_increment'

- **HUMIDITY_AVG (Average Relative Humidity):** This column was calculated by averaging the values from the "RH1" and "RH2" columns (presumably representing relative humidity readings from two different sources). This provides a single value for average relative humidity for each observation.

  'HUMIDITY_AVG','float','NO','',NULL,''

- **TEMPERATURE_DIFFERENCE (MAXT - MINT):** This column captures the difference between the maximum temperature (MAXT) and minimum temperature (MINT) for each observation. This can be a helpful variable for exploring how the range of temperature fluctuations might influence pest activity.

  'TEMPERATURE_DIFFERANCE','float','NO','',NULL,''

  'TEMP_DIFF','float','NO','',NULL,''

By implementing these feature engineering steps, the pest data was transformed into a more comprehensive and informative format, facilitating a deeper investigation of potential relationships between weather variables and pest populations across various locations and seasons in India.

Data Cleaning after transformations :

I have found Zero Null values among all the columns including the transformed columns.

Exploratory Data Analysis:

Data Overview:

- **A sample of the first 10 rows can be retrieved using**

  `SELECT * FROM PESTDATA LIMIT 10;`**.**

- **The total number of observations is** `17636`**, obtained using**

  `SELECT COUNT(*) FROM PESTDATA;`**.**

- **The data spans** `48 years` **based on the distinct observation years**

  (`SELECT count(distinct(OBSERVATIONYEAR)) FROM PESTDATA;`**).**

- **The distinct pest names can be identified using**

  `SELECT distinct(PESTNAME) FROM PESTDATA;`**.**

  'Gallmidge'

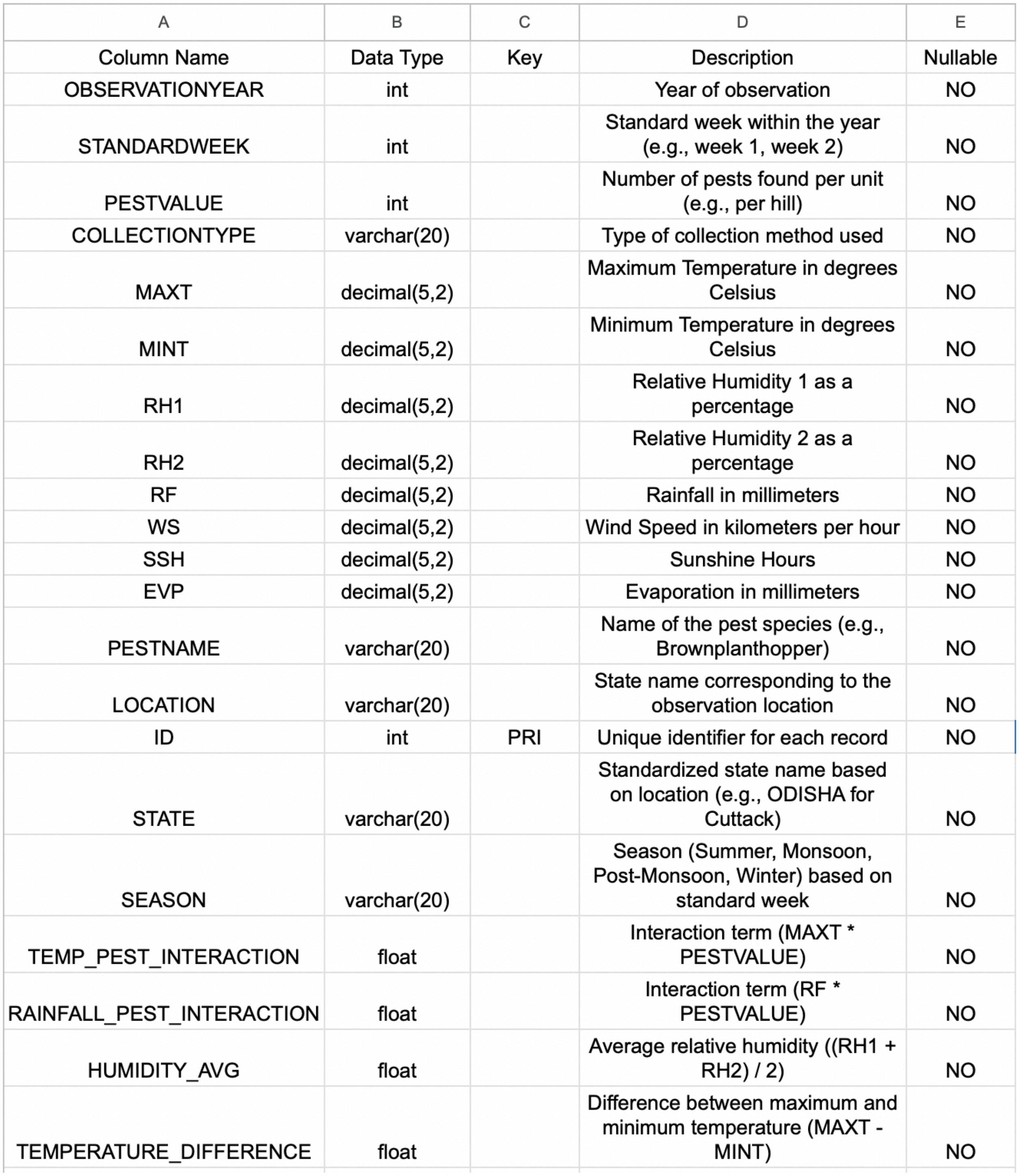
```
'Brownplanthopper'

'Greenleafhopper'

'LeafFolder'

'Yellowstemborer'

'Caseworm'

'Miridbug'

'ZigZagleafhopper'

'LeafBlast'

'NeckBlast'
```

- **The script queries the number of pest species found in each location using**

  `SELECT LOCATION, COUNT(PESTNAME) FROM PESTDATA GROUP BY LOCATION;`**.**

```
'Cuttack','1144'

'Ludhiana','1508'

'Maruteru','6169'

'Palampur','1248'

'Raipur','2028'

'Rajendranagar','5539'
```

Information on Data:



A	B	C	D	E
Column Name	Data Type	Key	Description	Nullable
OBSERVATIONYEAR	int		Year of observation	NO
STANDARDWEEK	int		Standard week within the year (e.g., week 1, week 2)	NO
PESTVALUE	int		Number of pests found per unit (e.g., per hill)	NO
COLLECTIONTYPE	varchar(20)		Type of collection method used	NO
MAXT	decimal(5,2)		Maximum Temperature in degrees Celsius	NO
MINT	decimal(5,2)		Minimum Temperature in degrees Celsius	NO
RH1	decimal(5,2)		Relative Humidity 1 as a percentage	NO
RH2	decimal(5,2)		Relative Humidity 2 as a percentage	NO
RF	decimal(5,2)		Rainfall in millimeters	NO
WS	decimal(5,2)		Wind Speed in kilometers per hour	NO
SSH	decimal(5,2)		Sunshine Hours	NO
EVP	decimal(5,2)		Evaporation in millimeters	NO
PESTNAME	varchar(20)		Name of the pest species (e.g., Brownplanthopper)	NO
LOCATION	varchar(20)		State name corresponding to the observation location	NO
ID	int	PRI	Unique identifier for each record	NO
STATE	varchar(20)		Standardized state name based on location (e.g., ODISHA for Cuttack)	NO
SEASON	varchar(20)		Season (Summer, Monsoon, Post-Monsoon, Winter) based on standard week	NO
TEMP_PEST_INTERACTION	float		Interaction term (MAXT * PESTVALUE)	NO
RAINFALL_PEST_INTERACTION	float		Interaction term (RF * PESTVALUE)	NO
HUMIDITY_AVG	float		Average relative humidity ((RH1 + RH2) / 2)	NO
TEMPERATURE_DIFFERENCE	float		Difference between maximum and minimum temperature (MAXT - MINT)	NO

QUESTIONS ON Basic Descriptive Statistics

- **WHAT ARE THE Total Number of Records?**

  '17636'

- **HOW MANY LOCATIONS OR STATE THE DATA IS CONSIDERED IN THE DATA?**

  'ANDHRA PRADESH'

  'CHHATTISGARH'

  'Himachal Pradesh'

  'MANIPUR'

  'ODISHA'

  'PUNJAB'

- **HOW MANY NUMBERS OF YEARS THE DATA IS CONSIDERED**

  '48'

- **WHAT ARE THE SEASONS IN THE DATA?**

  'MONSOON'

  'WINTER'

  'SUMMER'

  'POST-MONSOON'

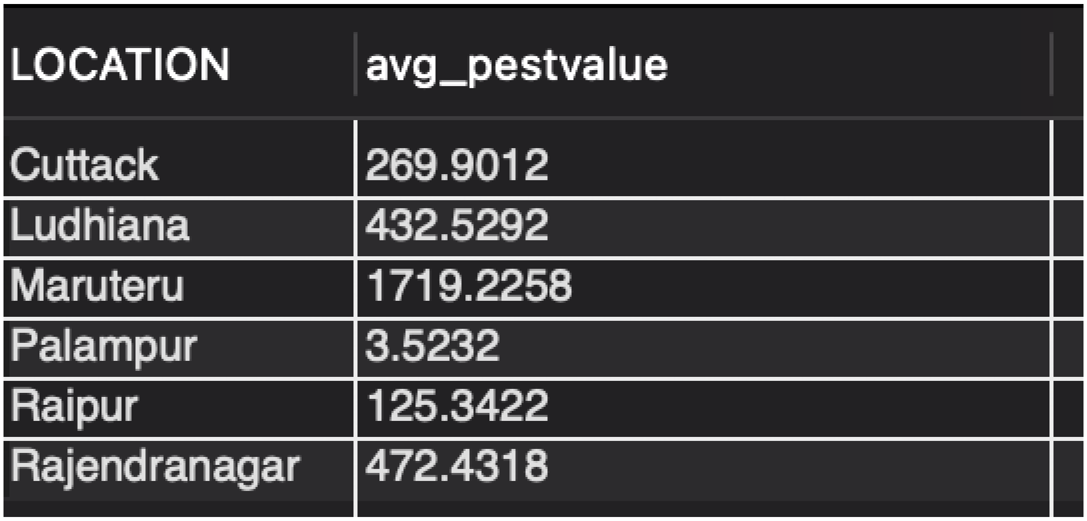
- **SUMMARY SATISTISTICS FOR TEMPERATURE WITH EACH STATE**

MAX	Min	State
'32.136976',	'22.008829',	'ODISHA'
'29.977188',	'17.436936',	'PUNJAB'
'31.060204',	'22.633993',	'ANDHRA PRADESH'
'23.764744',	'13.511458',	'Himachal Pradesh'
'32.920759',	'20.033432',	'CHHATTISGARH'
'32.479906',	'19.930872',	'MANIPUR'

- **Distribution of Pest Values Across Different Locations**

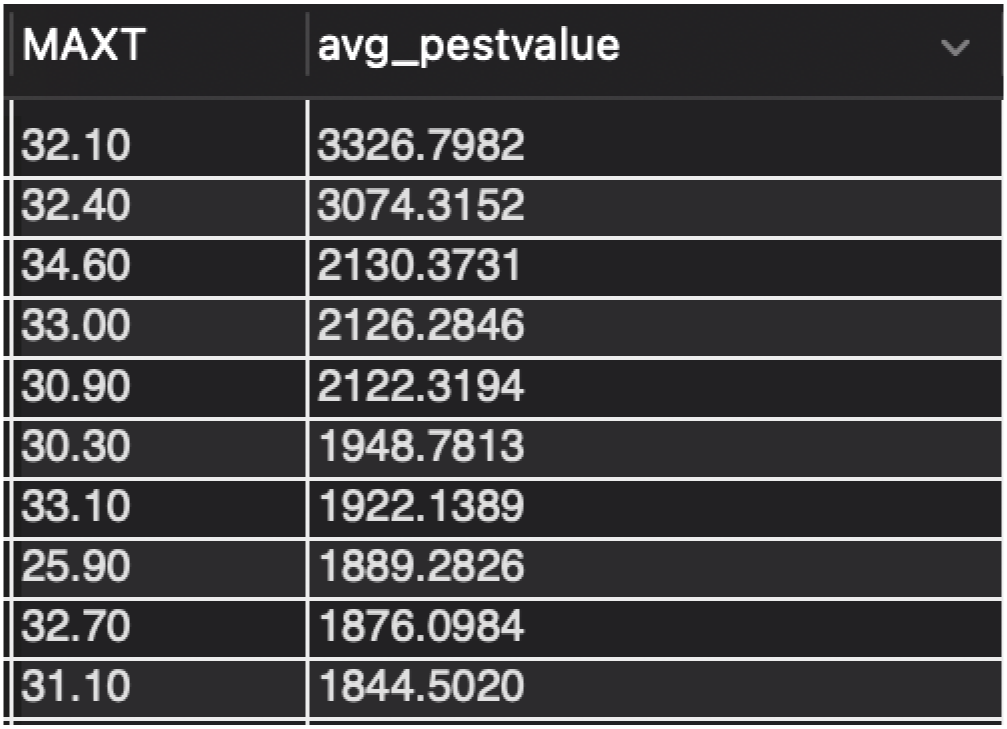
'Cuttack',	'269.9012'
'Ludhiana',	'432.5292'
'Maruteru',	'1719.2258'
'Palampur',	'3.5232'
'Raipur',	'125.3422'
'Rajendranagar',	'472.4318'

- **Question: How does the average pest value vary across different locations in India?**



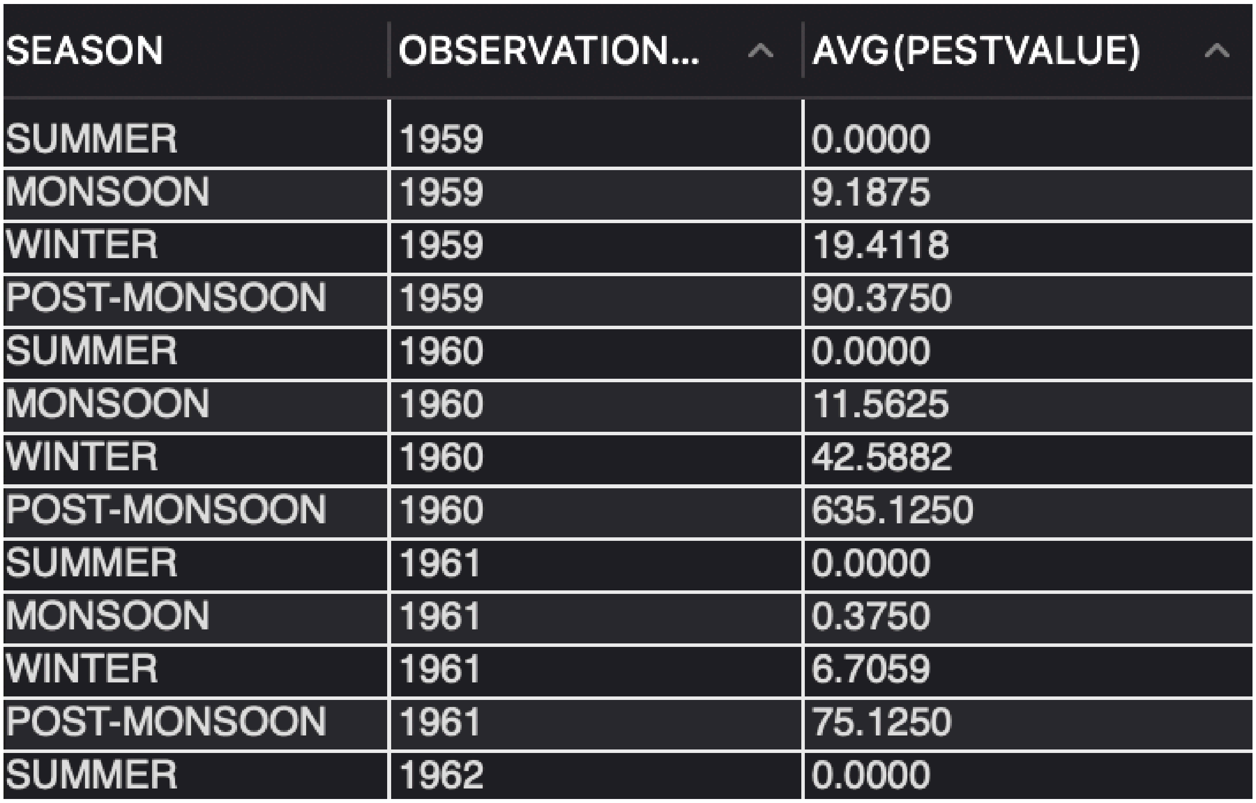
LOCATION	avg_pestvalue
Cuttack	269.9012
Ludhiana	432.5292
Maruteru	1719.2258
Palampur	3.5232
Raipur	125.3422
Rajendranagar	472.4318

- **Question: Is there a relationship between maximum temperature (MAXT) and average pest value?**



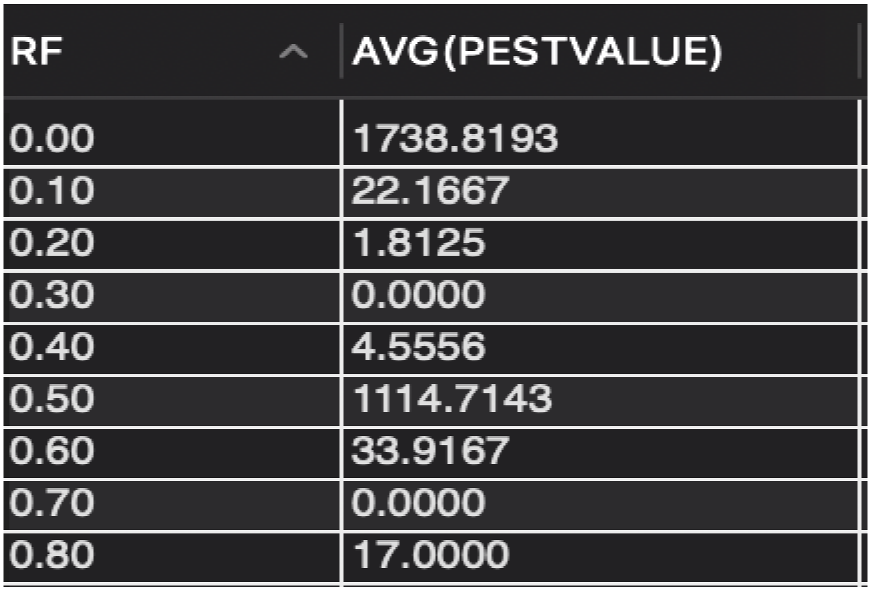
MAXT	avg_pestvalue
32.10	3326.7982
32.40	3074.3152
34.60	2130.3731
33.00	2126.2846
30.90	2122.3194
30.30	1948.7813
33.10	1922.1389
25.90	1889.2826
32.70	1876.0984
31.10	1844.5020

- **Question: How does the average pest value vary across different seasons (Summer, Monsoon, Post-Monsoon, Winter) for each year?**



SEASON	OBSERVATION...	AVG(PESTVALUE)
SUMMER	1959	0.0000
MONSOON	1959	9.1875
WINTER	1959	19.4118
POST-MONSOON	1959	90.3750
SUMMER	1960	0.0000
MONSOON	1960	11.5625
WINTER	1960	42.5882
POST-MONSOON	1960	635.1250
SUMMER	1961	0.0000
MONSOON	1961	0.3750
WINTER	1961	6.7059
POST-MONSOON	1961	75.1250
SUMMER	1962	0.0000

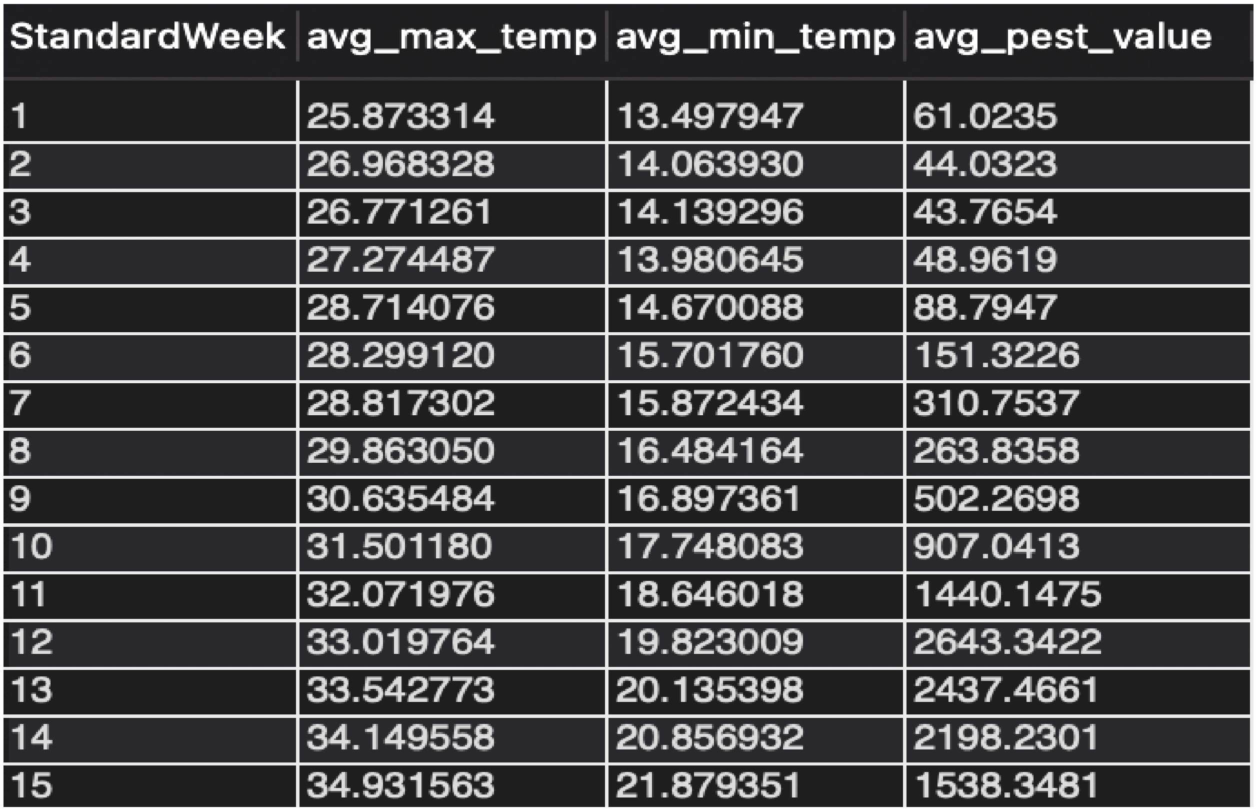
- **Question: How does rainfall (RF) affect average pest value during the summer season?**



RF	AVG(PESTVALUE)
0.00	1738.8193
0.10	22.1667
0.20	1.8125
0.30	0.0000
0.40	4.5556
0.50	1114.7143
0.60	33.9167
0.70	0.0000
0.80	17.0000

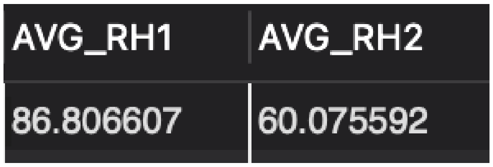
- **Question: How do average maximum temperature (avg_max_temp), average minimum temperature (avg_min_temp),and average pest value (avg_pest_value)**

vary across standard weeks throughout the year?



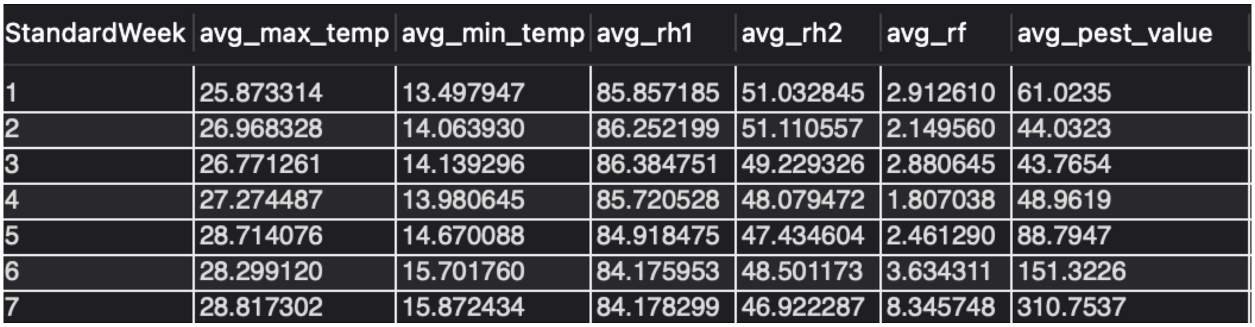
StandardWeek	avg_max_temp	avg_min_temp	avg_pest_value
1	25.873314	13.497947	61.0235
2	26.968328	14.063930	44.0323
3	26.771261	14.139296	43.7654
4	27.274487	13.980645	48.9619
5	28.714076	14.670088	88.7947
6	28.299120	15.701760	151.3226
7	28.817302	15.872434	310.7537
8	29.863050	16.484164	263.8358
9	30.635484	16.897361	502.2698
10	31.501180	17.748083	907.0413
11	32.071976	18.646018	1440.1475
12	33.019764	19.823009	2643.3422
13	33.542773	20.135398	2437.4661
14	34.149558	20.856932	2198.2301
15	34.931563	21.879351	1538.3481

- **Question: What are the average relative humidity levels (RH1 and RH2) for observations with pest values above the average?**



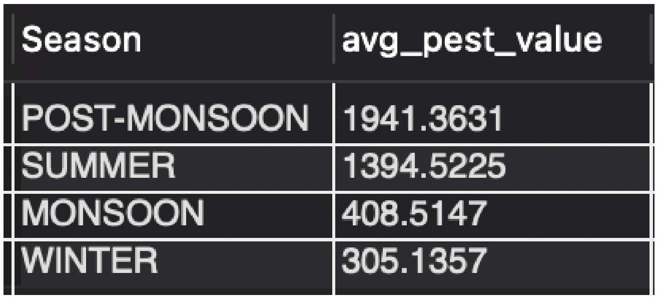
AVG_RH1	AVG_RH2
86.806607	60.075592

- **Question: How do various weather variables (average maximum temperature, average minimum temperature, average relative humidity, rainfall) and average pest value change across standard weeks throughout the year?**



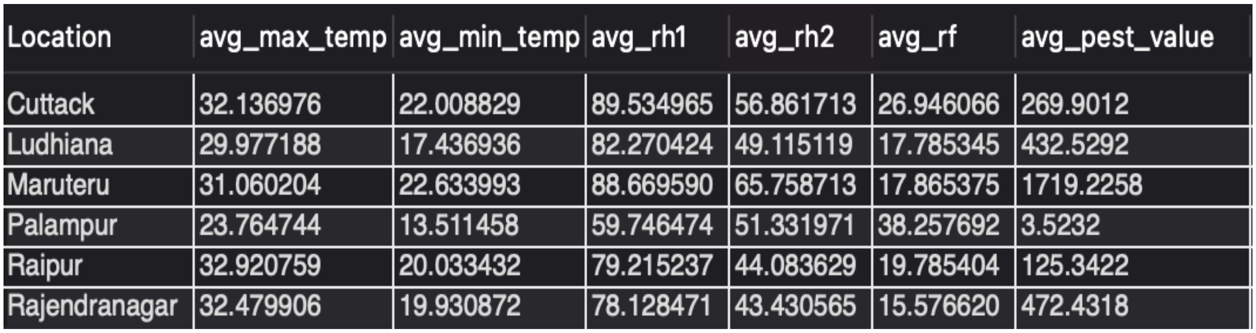
StandardWeek	avg_max_temp	avg_min_temp	avg_rh1	avg_rh2	avg_rf	avg_pest_value
1	25.873314	13.497947	85.857185	51.032845	2.912610	61.0235
2	26.968328	14.063930	86.252199	51.110557	2.149560	44.0323
3	26.771261	14.139296	86.384751	49.229326	2.880645	43.7654
4	27.274487	13.980645	85.720528	48.079472	1.807038	48.9619
5	28.714076	14.670088	84.918475	47.434604	2.461290	88.7947
6	28.299120	15.701760	84.175953	48.501173	3.634311	151.3226
7	28.817302	15.872434	84.178299	46.922287	8.345748	310.7537

- **Question: Which season (Summer, Monsoon, Post-Monsoon, Winter) has the highest average pest value?**



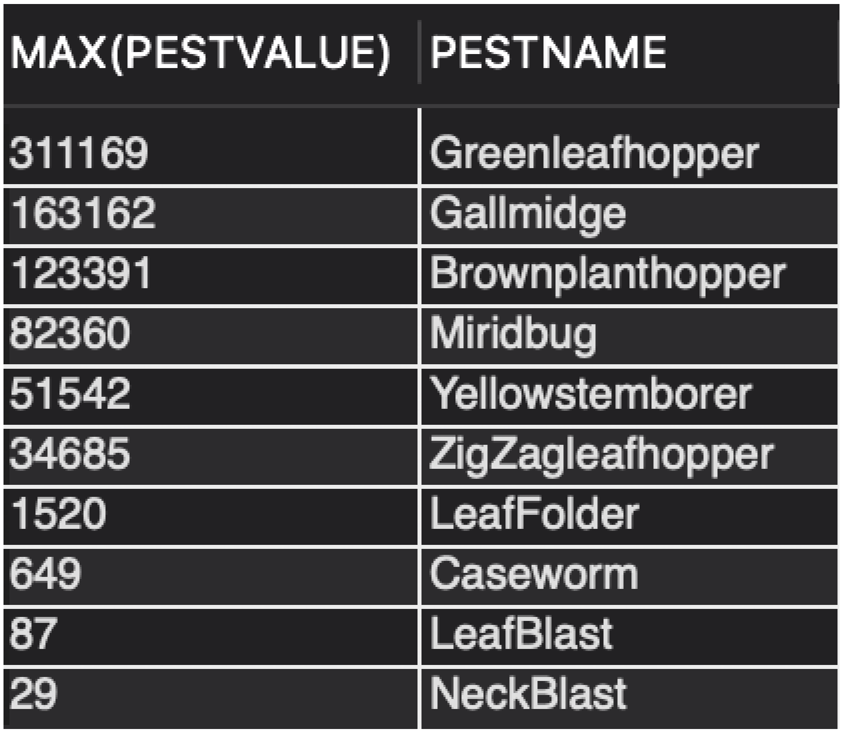
Season	avg_pest_value
POST-MONSOON	1941.3631
SUMMER	1394.5225
MONSOON	408.5147
WINTER	305.1357

- **Question: How do average weather variables (maximum temperature, minimum temperature, relative humidity, rainfall) and average pest value vary across different locations?**



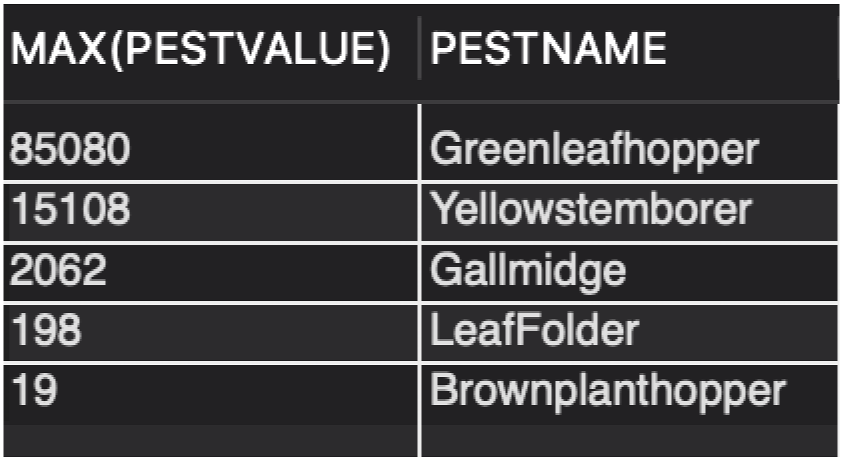
Location	avg_max_temp	avg_min_temp	avg_rh1	avg_rh2	avg_rf	avg_pest_value
Cuttack	32.136976	22.008829	89.534965	56.861713	26.946066	269.9012
Ludhiana	29.977188	17.436936	82.270424	49.115119	17.785345	432.5292
Maruteru	31.060204	22.633993	88.669590	65.758713	17.865375	1719.2258
Palampur	23.764744	13.511458	59.746474	51.331971	38.257692	3.5232
Raipur	32.920759	20.033432	79.215237	44.083629	19.785404	125.3422
Rajendranagar	32.479906	19.930872	78.128471	43.430565	15.576620	472.4318

- **Question: Which pest species has the highest maximum pest value across all states?**



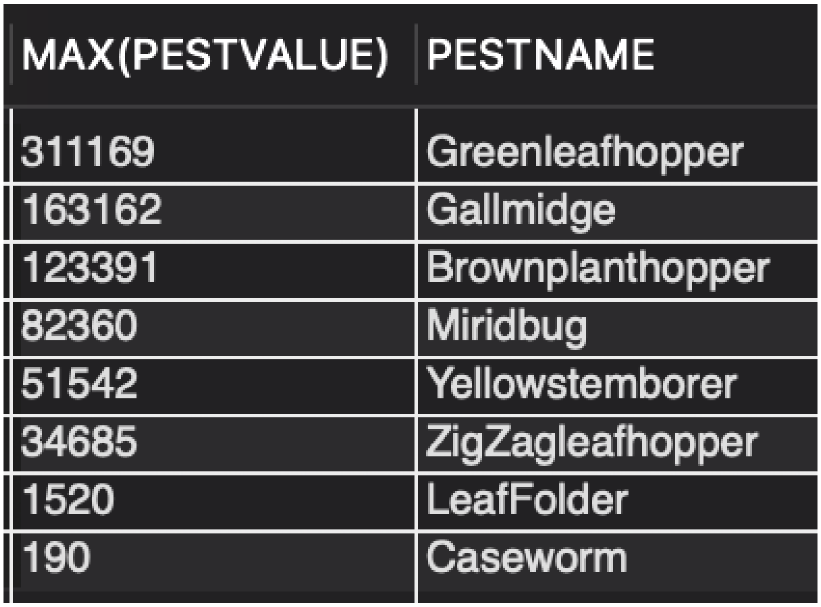
MAX(PESTVALUE)	PESTNAME
311169	Greenleafhopper
163162	Gallmidge
123391	Brownplanthopper
82360	Miridbug
51542	Yellowstemborer
34685	ZigZagleafhopper
1520	LeafFolder
649	Caseworm
87	LeafBlast
29	NeckBlast

- **Question (for each location): (Replace `[Location Name]` with the specific location, like Cuttack, Maruteru, etc.)**
  Cuttack



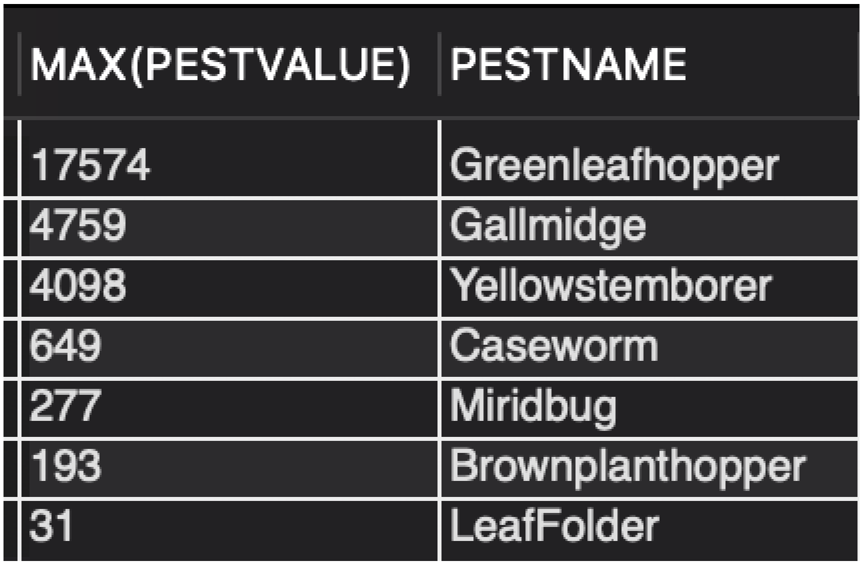
MAX(PESTVALUE)	PESTNAME
85080	Greenleafhopper
15108	Yellowstemborer
2062	Gallmidge
198	LeafFolder
19	Brownplanthopper

Maruteru



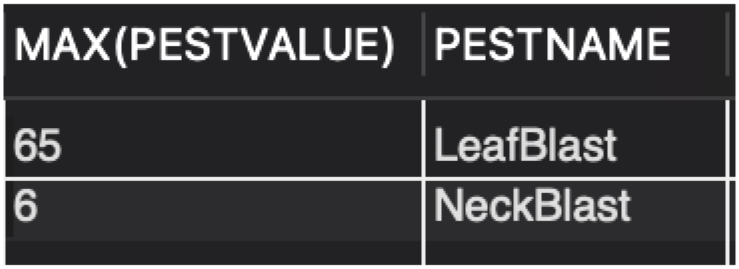
MAX(PESTVALUE)	PESTNAME
311169	Greenleafhopper
163162	Gallmidge
123391	Brownplanthopper
82360	Miridbug
51542	Yellowstemborer
34685	ZigZagleafhopper
1520	LeafFolder
190	Caseworm

Raipur



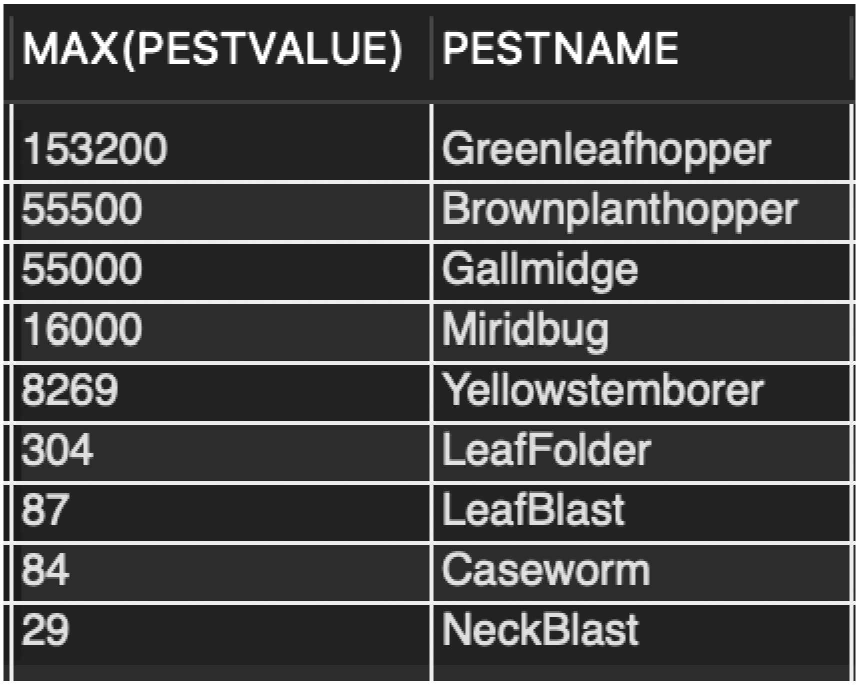
MAX(PESTVALUE)	PESTNAME
17574	Greenleafhopper
4759	Gallmidge
4098	Yellowstemborer
649	Caseworm
277	Miridbug
193	Brownplanthopper
31	LeafFolder

Palampur



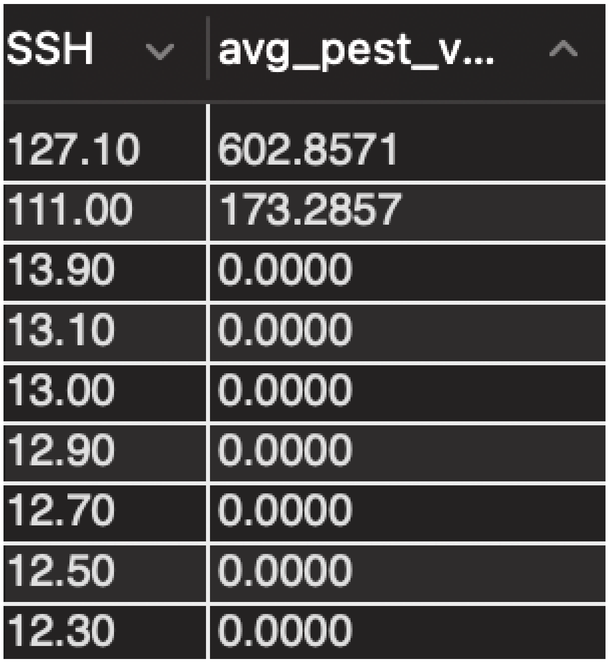
MAX(PESTVALUE)	PESTNAME
65	LeafBlast
6	NeckBlast

Rajendranagar

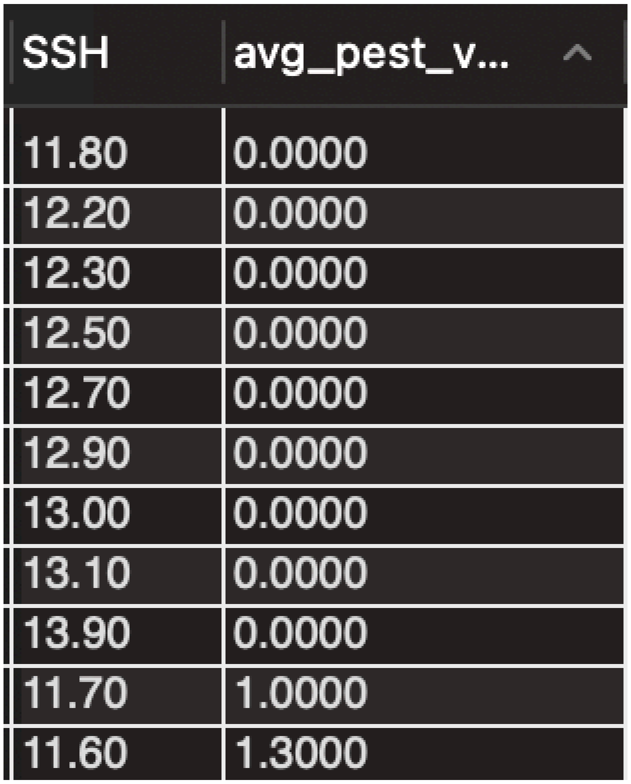


MAX(PESTVALUE)	PESTNAME
153200	Greenleafhopper
55500	Brownplanthopper
55000	Gallmidge
16000	Miridbug
8269	Yellowstemborer
304	LeafFolder
87	LeafBlast
84	Caseworm
29	NeckBlast

- **Question: Is there a relationship between sunshine hours (SSH) and average pest value?**



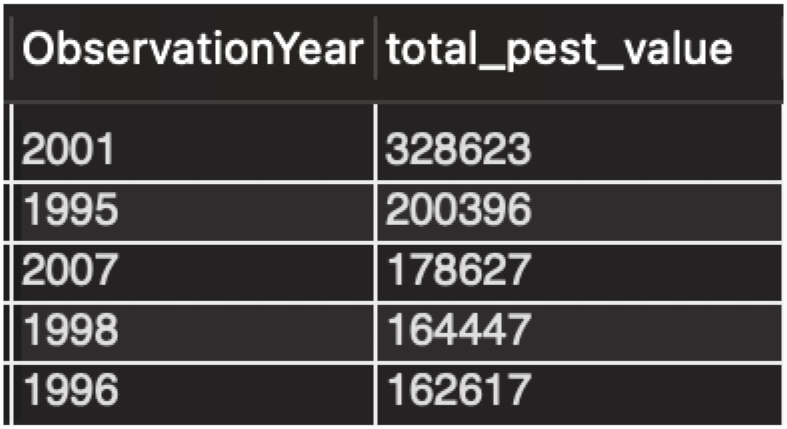
SSH	avg_pest_v...
127.10	602.8571
111.00	173.2857
13.90	0.0000
13.10	0.0000
13.00	0.0000
12.90	0.0000
12.70	0.0000
12.50	0.0000
12.30	0.0000



SSH	avg_pest_v...
11.80	0.0000
12.20	0.0000
12.30	0.0000
12.50	0.0000
12.70	0.0000
12.90	0.0000
13.00	0.0000
13.10	0.0000
13.90	0.0000
11.70	1.0000
11.60	1.3000

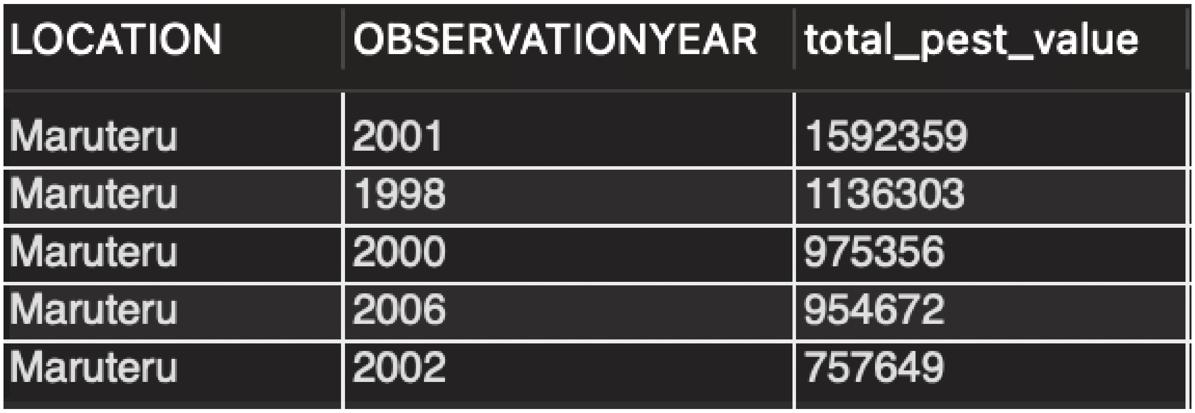
While the provided data snippet doesn't allow for a definitive conclusion, we can explore the relationship between sunshine hours (SSH) and average pest value further Incorporating factors like temperature, rainfall, season, and pest species might reveal more nuanced relationships. Also creating a scatter plot with SSH on the x-axis and average pest value on the y-axis can provide a visual indication of any correlation.

- **Question: Which five years have the highest total pest value summed across all standard weeks?**



ObservationYear	total_pest_value
2001	328623
1995	200396
2007	178627
1998	164447
1996	162617

- **Question: Which five locations have the highest total pest value summed across all observation years?**



LOCATION	OBSERVATIONYEAR	total_pest_value
Maruteru	2001	1592359
Maruteru	1998	1136303
Maruteru	2000	975356
Maruteru	2006	954672
Maruteru	2002	757649