Visual Exploration of Dog Intelligence Relative to Size

A MINI PROJECT REPORT

18CSE391T – BIG DATA TOOLS

AND TECHNIQUES

Submitted by

DANIA B SAM
[RA2111027010071.]

Under the guidance of

Dr.K.Sornalakshmi

Assistant Professor, Department of Computer Science and Engineering

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

of

FACULTY OF ENGINEERING AND TECHNOLOGY

S.R.M. Nagar, Kattankulathur, Chengalpattu District

MAY 2024

DATASET DESCRIPTION :-

The dataset provides information on dog intelligence comparison based on sizes, derived from research conducted by Stanley Coren, a professor of canine psychology at the University of British Columbia. Coren's work, initially published in 1994, has been a subject of debate but has gradually gained acceptance over time.

The dataset includes the following attributes:

- **Breed**: The breed of the dog.
- **height_low_inches**: The lower bound of the height range of the dog in inches.
- **height_high_inches**: The upper bound of the height range of the dog in inches.
- **weight_low_lbs**: The lower bound of the weight range of the dog in pounds.
- **weight_high_lbs**: The upper bound of the weight range of the dog in pounds.
- **reps_lower**: Lower bound of the number of repetitions of a command needed for the dog to learn it.
- **reps_upper**: Upper bound of the number of repetitions of a command needed for the dog to learn it.

The dataset contains 150 entries, each representing a different breed of dog. The "reps_lower" and "reps_upper" columns have some missing values (NaNs) in 14 entries. This dataset provides valuable insights into how Dog intelligence relates to size, enabling deeper analysis and visualization of these connections.

	Breed	height_low_inches	height_high_inches	weight_low_lbs	weight_high_lbs	reps_lower	reps_upper
0	Akita	26.0	28.0	80	120	1.0	4.0
1	A0tolian Sheepdog	27.0	29.0	100	150	1.0	4.0
2	Bernese Mountain Dog	23.0	27.0	85	110	1.0	4.0
3	Bloodhound	24.0	26.0	80	120	1.0	4.0
4	Borzoi	26.0	28.0	70	100	1.0	4.0
...	...	...	...	...	...	...	...
145	Papillon	8.0	11.0	5	10	NaN	NaN
146	Pomeranian	12.0	12.0	3	7	NaN	NaN
147	Poodle Toy	10.0	10.0	10	10	NaN	NaN
148	Toy Fox Terrier	10.0	10.0	4	7	NaN	NaN
149	Yorkshire Terrier	8.0	8.0	3	7	NaN	NaN

150 rows × 7 columns

To address the issue of missing values in the "reps_lower" and "reps_upper" columns, the fillna() method was employed to replace them with the mean values of each corresponding column. This approach was crucial for maintaining the dataset's completeness and readiness for visualization. By substituting null values with the mean, the dataset's integrity was upheld, ensuring that missing data did not hinder the visualization process. This method facilitated a more accurate analysis of the correlation between dog intelligence and size, enabling insightful conclusions to be drawn from the data.

```
df['reps_lower'].fillna(df['reps_lower'].mean(), inplace=True)
df['reps_upper'].fillna(df['reps_upper'].mean(), inplace=True)
```

```
missing_values = df.isnull().sum()
missing_values
```

```
Breed                0
height_low_inches    0
height_high_inches   0
weight_low_lbs       0
weight_high_lbs      0
reps_lower           0
reps_upper           0
dtype: int64
```

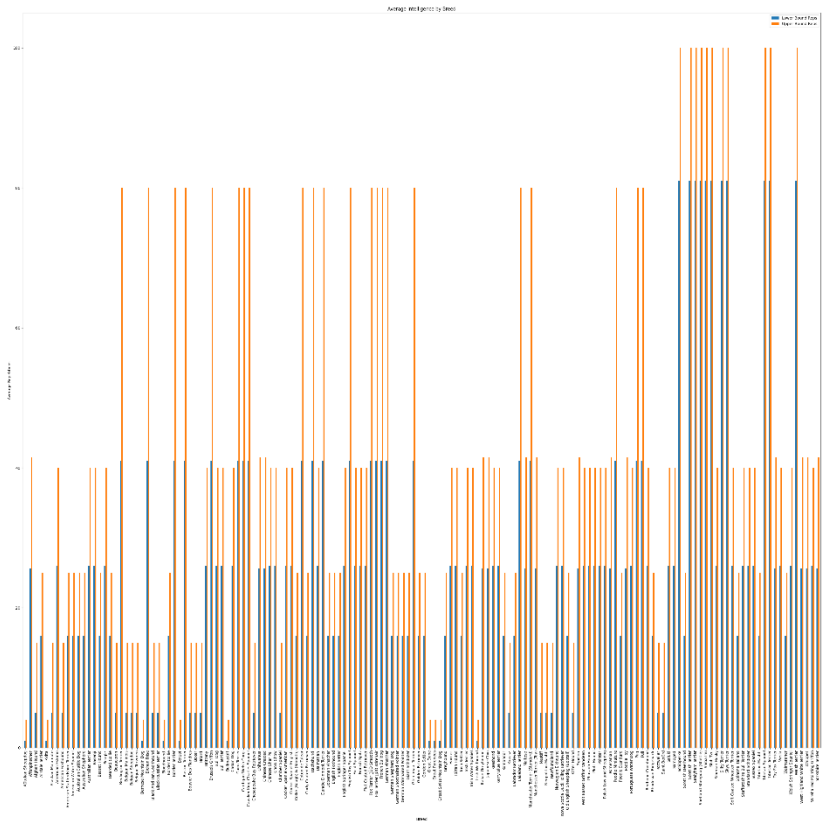
1. Bar Plot

Code Snippet :-

```
intelligence_by_breed = df.groupby('Breed')[['reps_lower', 'reps_upper']].mean()

# Plot the bar graph
intelligence_by_breed.plot(kind='bar', figsize=(30, 30))
plt.title('Average Intelligence by Breed')
plt.xlabel('Breed')
plt.ylabel('Average Repetitions')
plt.xticks(rotation=90)
plt.legend(['Lower Bound Reps', 'Upper Bound Reps'])
plt.tight_layout()
plt.show()
```

```
# Find the breed with the highe Loading... repetitions
highest_breed = intelligence_by_breed['reps_lower'].idxmax()
highest_reps = intelligence_by_breed['reps_lower'].max()

# Find the breed with the lowest average repetitions
lowest_breed = intelligence_by_breed['reps_lower'].idxmin()
lowest_reps = intelligence_by_breed['reps_lower'].min()

print(f"The breed with the highest average repetitions is '{highest_breed}' with {highest_reps} repetitions.")
print(f"The breed with the lowest average repetitions is '{lowest_breed}' with {lowest_reps} repetitions.")
```

Screenshot:-



Inference :-

The Bar Graph shows that how different dog breeds grasp new commands, showing us the average number of repetitions they need to learn. In the graph, the bright orange bars represent the upper limit of repetitions, while the blue show the lower limit. At the top of the list is the 'Schipperke' breed, needing an average of 81.0 repetitions, while the 'A0tolian Sheepdog' requires just 1.0 repetitions, making it the quickest learner. Based on the analysis, breeds with averages above 95% are like the brightest dogs, with less than 5 repetitions. Those between 85% and 95% are considered as excellent workers, and anything above 70% indicates above average learners. Dogs in the 50% to 70% range are average learners, while those below 30% might need a little extra patience during training.

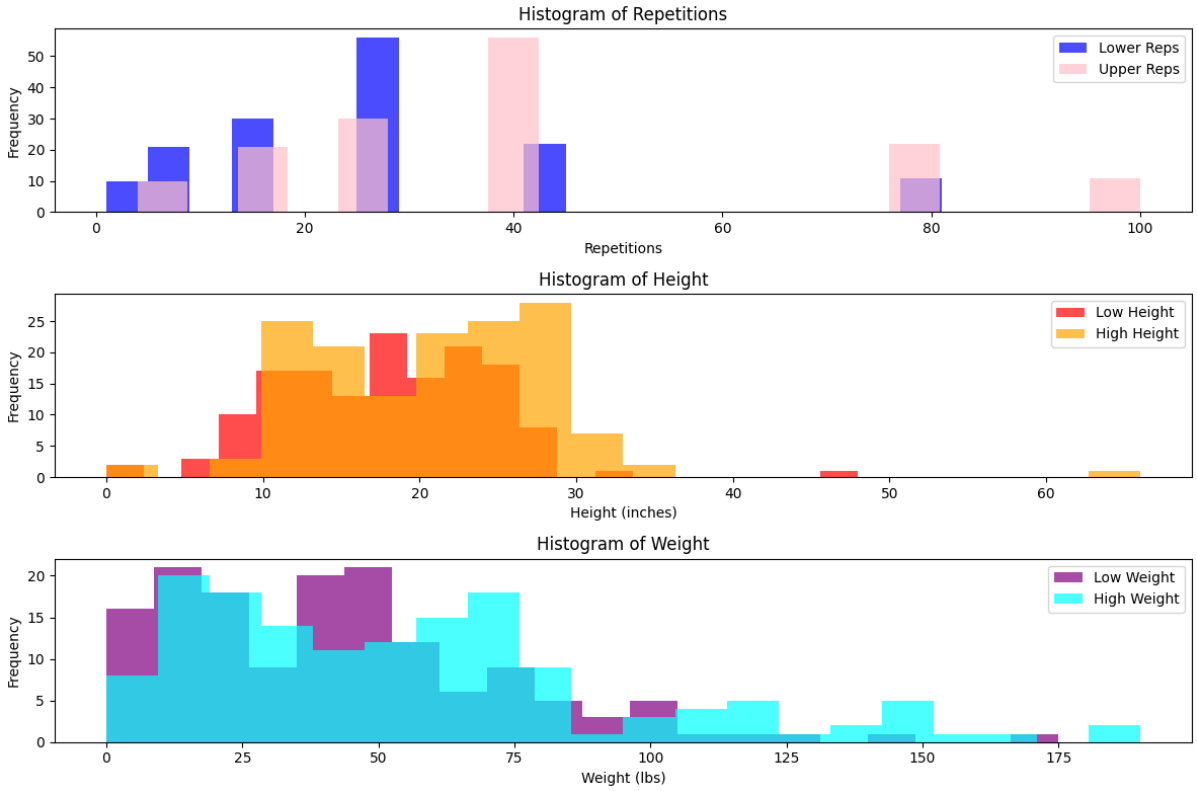
2. Histogram

Code Snippet :-

```
plt.figure(figsize=(12, 8))

# Histogram for repetitions
plt.subplot(3, 1, 1)
plt.hist(df['reps_lower'], bins=20, color='blue', alpha=0.7, label='Lower Reps')
plt.hist(df['reps_upper'], bins=20, color='pink', alpha=0.7, label='Upper Reps')
plt.title('Histogram of Repetitions')
plt.xlabel('Repetitions')
plt.ylabel('Frequency')
plt.legend()

# Histogram for height
plt.subplot(3, 1, 2)
plt.hist(df['height_low_inches'], bins=20, color='red', alpha=0.7, label='Low Height')
plt.hist(df['height_high_inches'], bins=20, color='orange', alpha=0.7, label='High Height')
plt.title('Histogram of Height')
plt.xlabel('Height (inches)')
plt.ylabel('Frequency')
plt.legend()
```

```
# Histogram for weight
plt.subplot(3, 1, 3)
plt.hist(df['weight_low_lbs'], bins=20, color='purple', alpha=0.7, label='Low Weight')
plt.hist(df['weight_high_lbs'], bins=20, color='cyan', alpha=0.7, label='High Weight')
plt.title('Histogram of Weight')
plt.xlabel('Weight (lbs)')
plt.ylabel('Frequency')
plt.legend()

plt.tight_layout()
plt.show()
```

Screenshot :-



Inference :-

Repetitions: The histogram shows that a lot of dog breeds seem to learn commands after around 30 repetitions, and there is another group needing about 40 repetitions, which is quite common too. This hints at a bunch of breeds with similar learning skills, sort of grouped around these repetition numbers.

Height: The histogram illustrates that the frequency of dogs with higher

heights is more, particularly in the range of 20-30 inches. However, there is significant overlap between dogs of low and high heights, indicating no significant difference between them.

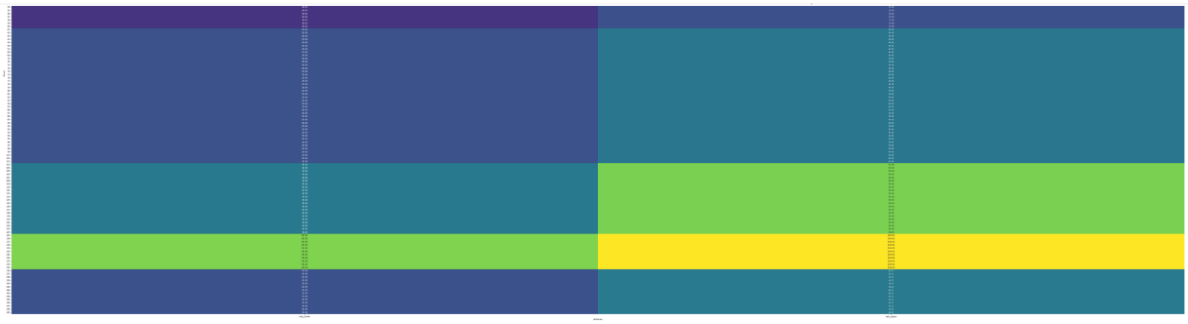
Weight: Similarly, the histogram for weight shows overlap between dogs with low and high weights, suggesting no significant difference in the frequency distribution based on weight categories.

3. HEAT MAP

Code Snippet :-

```
plt.figure(figsize=(100, 100))

# Heatmap for repetitions
plt.subplot(3, 1, 1)
sns.heatmap(df[['reps_lower', 'reps_upper']], cmap='viridis', annot=True, fmt=".2f")
plt.title('Heatmap of Repetitions')
plt.xlabel('Attribute')
plt.ylabel('Breed')


plt.tight_layout()
plt.show()
```

Screenshot:-



Inference :-

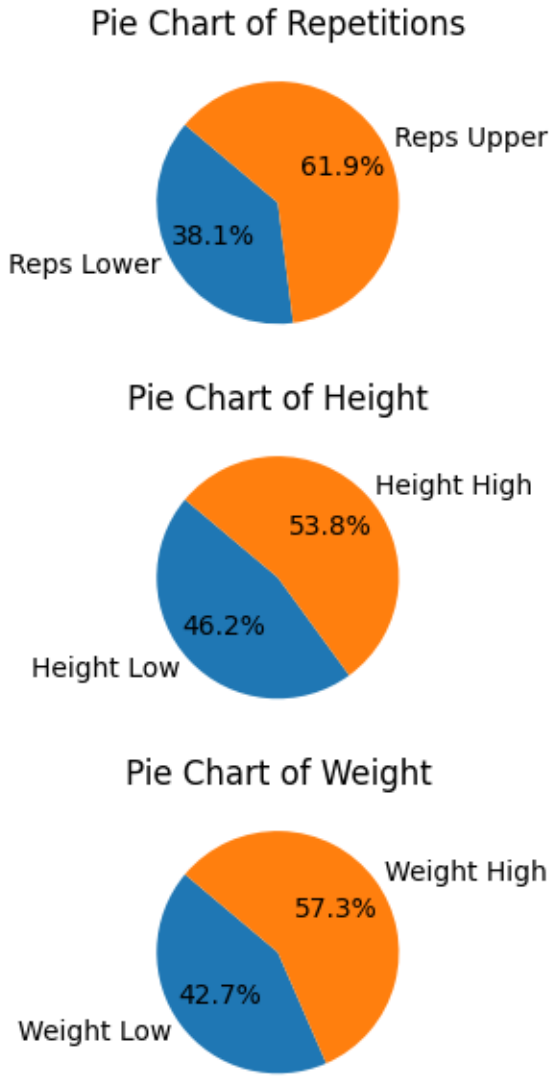
The heatmap analysis using the Viridis colormap reveals patterns in the repetition of commands across different breeds. This analysis suggests varying levels of command repetition requirements across different breeds. The Breed ranked from 1-10 required fewer repetitions which was indicated by a deep blue colour . The Breeds ranked from 125-135 required a need for high repetition which was indicated by a vibrant yellow-green hues. Breeds listed between 61-124 exhibit a gradual increase in need of repetition of commands.

4. PIE CHART

Code Snippet :-

```
plt.figure(figsize=(10,6))

# Pie chart for repetitions
plt.subplot(3, 1, 1)
reps_counts = df[['reps_lower', 'reps_upper']].sum().tolist()
plt.pie(reps_counts, labels=['Reps Lower', 'Reps Upper'], autopct='%1.1f%%', startangle=140)
plt.title('Pie Chart of Repetitions')

# Pie chart for height
plt.subplot(3, 1, 2)
height_counts = df[['height_low_inches', 'height_high_inches']].sum().tolist()
plt.pie(height_counts, labels=['Height Low', 'Height High'], autopct='%1.1f%%', startangle=140)
plt.title('Pie Chart of Height')

# Pie chart for weight
plt.subplot(3, 1, 3)
weight_counts = df[['weight_low_lbs', 'weight_high_lbs']].sum().tolist()
plt.pie(weight_counts, labels=['Weight Low', 'Weight High'], autopct='%1.1f%%', startangle=140)
plt.title('Pie Chart of Weight')

plt.tight_layout()
plt.show()
```

Screenshot:-



Inference :-

The pie chart analysis of dog intelligence presents a breakdown on upper and lower bounds across various paraments . In the repetition chart , the chart indicates that 61.9% of the intelligence distributions corresponds to dog exhibiting a higher repetition level, depicted by the orange segment. 38.1% of the intelligence distribution is attributed to dogs with a lower repetition level, represented by the blue segment. In the height chart , the analysis reveals that 53.8% of dogs in the intelligence distribution is among

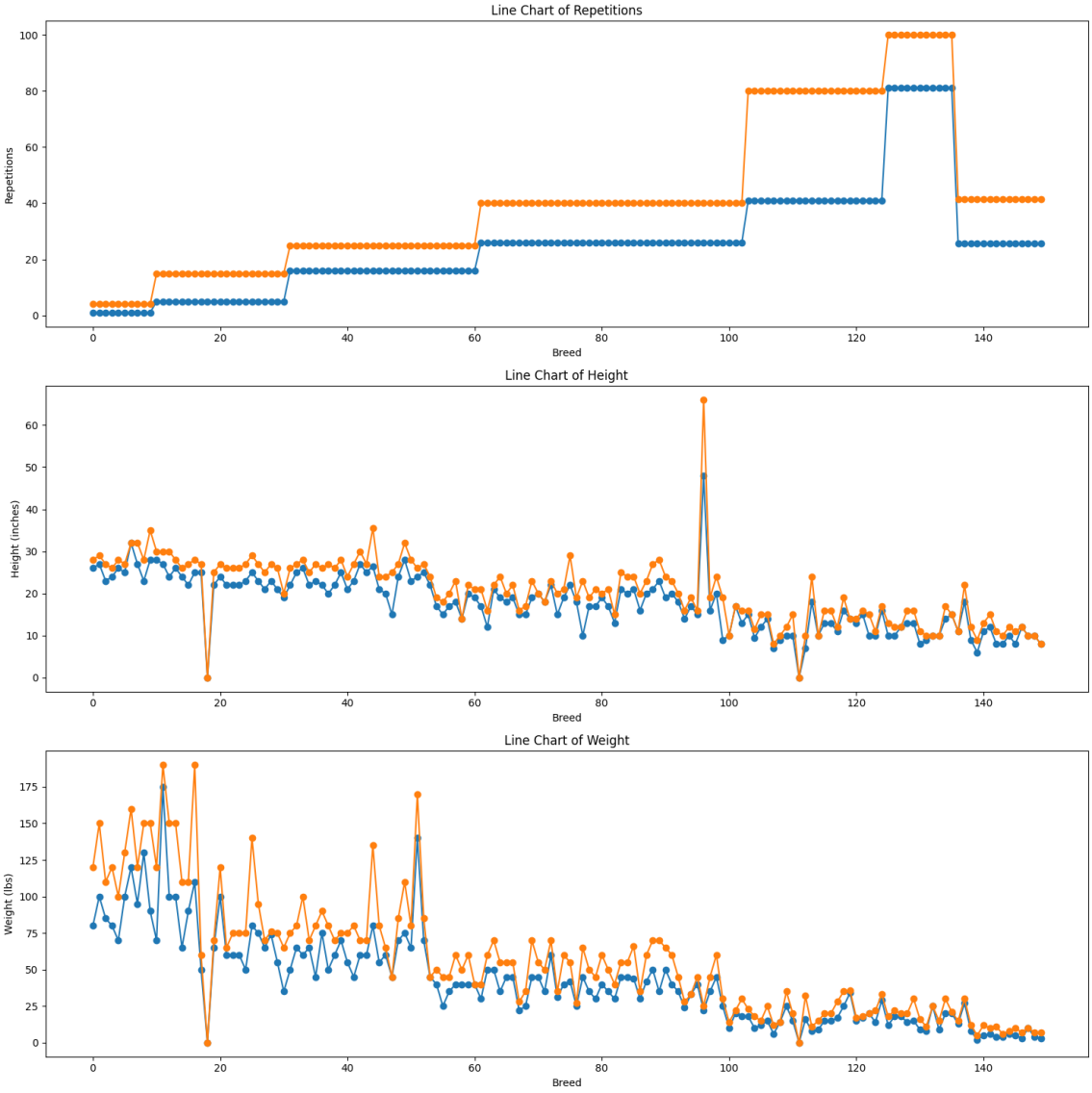
higher heights and 46.2% of dogs display lower heights in the intelligence distribution. In the weight chart, pie chart demonstrates that 57.3% of intelligent dogs possess higher weights and 42.7% of intelligent dogs have lower weights.

5. Line Chart

Code Snippet :-

```
plt.figure(figsize=(15,15))

# Line chart for repetitions
plt.subplot(3, 1, 1)
plt.plot(df[['reps_lower', 'reps_upper']], marker='o')
plt.title('Line Chart of Repetitions')
plt.xlabel('Breed')
plt.ylabel('Repetitions')

# Line chart for height
plt.subplot(3, 1, 2)
plt.plot(df[['height_low_inches', 'height_high_inches']], marker='o')
plt.title('Line Chart of Height')
plt.xlabel('Breed')
plt.ylabel('Height (inches)')

# Line chart for weight
plt.subplot(3, 1, 3)
plt.plot(df[['weight_low_lbs', 'weight_high_lbs']], marker='o')
plt.title('Line Chart of Weight')
plt.xlabel('Breed')
plt.ylabel('Weight (lbs)')

plt.tight_layout()
plt.show()
```

Screenshot :-



Inference :-

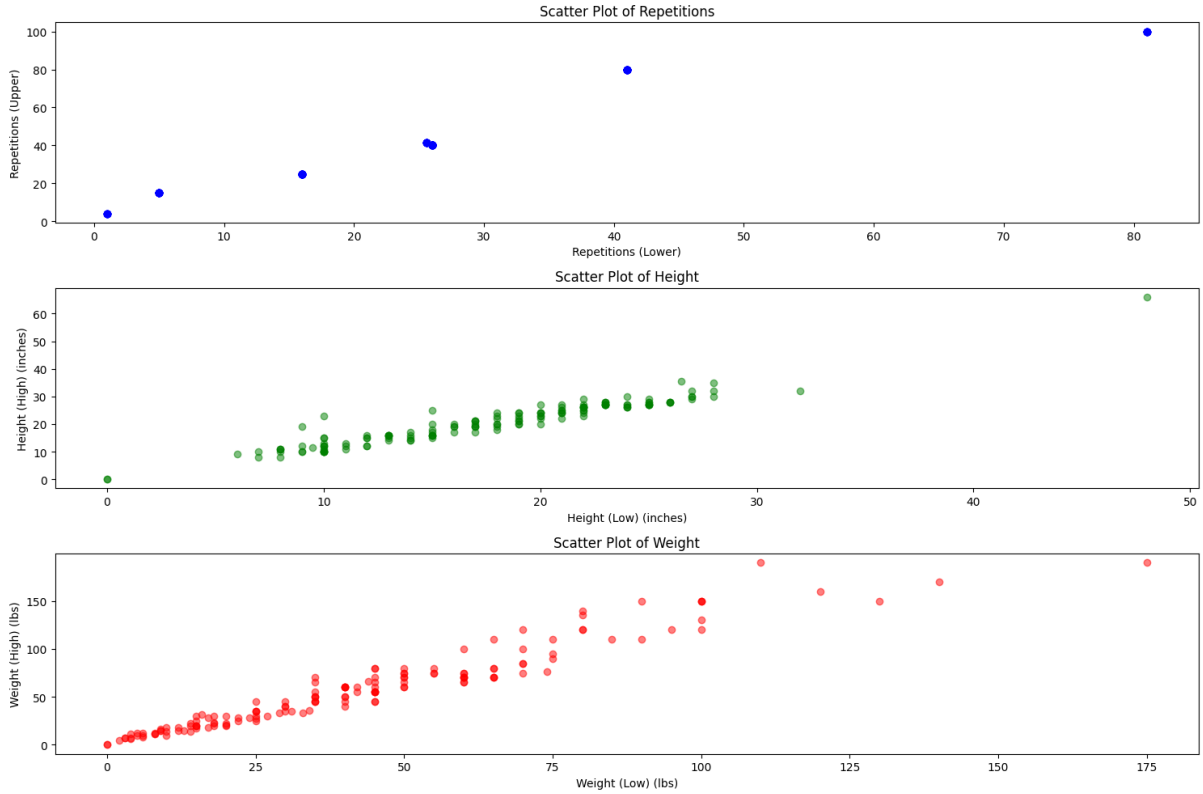
The line chart analysis reveals dynamic trends in dog intelligence across breeds. Repetition levels show fluctuations, starting and ending steadily but peaking around the 130th breed before sharply declining. Height exhibits significant changes, decreasing notably around the 20th breed, gradually rising and falling until a substantial increase after the 100th breed. Weight patterns fluctuate, with a gradual increase from the 10th to 20th breed, a sudden drop at the 20th, and subsequent fluctuations. These insights highlight how dog breeds differ in intelligence and physical traits like height and weight.

6. Scatter Plot

Code Snippet:-

```
plt.figure(figsize=(15,10))

# Scatter plot for repetitions
plt.subplot(3, 1, 1)
plt.scatter(df['reps_lower'], df['reps_upper'], marker='o', color='blue', alpha=0.5)
plt.title('Scatter Plot of Repetitions')
plt.xlabel('Repetitions (Lower)')
plt.ylabel('Repetitions (Upper)')

# Scatter plot for height
plt.subplot(3, 1, 2)
plt.scatter(df['height_low_inches'], df['height_high_inches'], marker='o', color='green', alpha=0.5)
plt.title('Scatter Plot of Height')
plt.xlabel('Height (Low) (inches)')
plt.ylabel('Height (High) (inches)')

# Scatter plot for weight
plt.subplot(3, 1, 3)
plt.scatter(df['weight_low_lbs'], df['weight_high_lbs'], marker='o', color='red', alpha=0.5)
plt.title('Scatter Plot of Weight')
plt.xlabel('Weight (Low) (lbs)')
plt.ylabel('Weight (High) (lbs)')

plt.tight_layout()
plt.show()
```

Screenshot:-



Inference :-

In the scatter plot analysis, there's a strong positive correlation between the higher and lower bounds of repetition (correlation coefficient: 0.94), height (correlation coefficient: 0.95), and weight (correlation coefficient: 0.96). This suggests that as the upper and lower bounds of repetition, height, and weight increase, they tend to do so together across the dataset.