MCSE ORANGE PROBLEM
SET-14 : Parkinson's Disease Progression

SRN:
1. PES2UG24CS157
2. PES2UG24CS160
3. PES2UG24CS162

```
import pandas as pd
df = pd.read_csv("./Set 14 Parkinsons Dataset.csv")
# first 5 rows:
df.head()
```

	Age	Sex	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	Jitter(%)	Shimmer	NHR	HNR	RPDE	DFA	Status
0	78.0	Male	140.121789	190.536786	97.447843	0.003266	0.019379	0.002041	16.498254	0.500525	0.533560	0.0
1	68.0	Male	135.891557	184.144247	146.516073	NaN	0.032288	0.013388	NaN	0.607867	0.588044	NaN
2	54.0	Male	137.926467	217.007154	146.633691	0.004332	0.026894	NaN	NaN	0.307454	0.552765	0.0
3	82.0	Male	137.125859	213.426320	104.141422	0.003044	0.031066	0.026576	19.716553	0.425107	0.620578	1.0
4	47.0	Female	135.797214	206.217093	NaN	0.002585	0.019409	0.024570	22.456411	NaN	0.664935	1.0

```
df.columns
```

```
Index(['Age', 'Sex', 'MDVP:Fo(Hz)', 'MDVP:Fhi(Hz)', 'MDVP:Flo(Hz)',
       'Jitter(%)', 'Shimmer', 'NHR', 'HNR', 'RPDE', 'DFA', 'Status'],
      dtype='object')
```

```
df.dtypes
```

```
Age              float64
Sex               object
MDVP:Fo(Hz)      float64
MDVP:Fhi(Hz)     float64
MDVP:Flo(Hz)     float64
Jitter(%)        float64
Shimmer          float64
NHR              float64
HNR              float64
RPDE             float64
DFA              float64
Status           float64
dtype: object
```

```
df.tail() #last 5 rows.
```

	Age	Sex	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	Jitter(%)	Shimmer	NHR	HNR	RPDE	DFA	Status
645	58.0	Female	141.534583	206.650626	140.518983	0.003342	0.028789	0.019267	18.269429	0.540974	0.586255	1.0
646	59.0	Male	175.959507	212.975999	146.602855	0.001406	0.031085	0.015520	20.017224	0.406982	0.615466	1.0
647	57.0	Male	172.922882	NaN	117.515492	0.004694	0.024181	0.009953	22.408343	0.439061	0.608174	1.0
648	80.0	Female	173.193216	225.242261	111.365171	0.004195	0.001892	0.030307	21.937337	0.527227	0.576940	1.0
649	53.0	Female	115.344607	173.836800	113.127531	0.003478	0.024831	0.013481	20.761187	0.410108	0.620830	0.0

```
Missing values per column:
 Age            33
Sex            30
MDVP:Fo(Hz)    35
MDVP:Fhi(Hz)   30
MDVP:Flo(Hz)   28
Jitter(%)      35
Shimmer        30
NHR            32
HNR            30
RPDE           26
DFA            29
Status         34
dtype: int64
Duplicates (found, removed): 0
```

Descriptive Statistics:

```
df['Age'].describe()
```

```
count    650.000000
mean      62.672609
std       12.605173
min       40.000000
25%       52.000000
50%       63.000000
75%       73.000000
max       84.000000
Name: Age, dtype: float64
```

```
df['MDVP:Fo(Hz)'].describe()
```

```
count    650.000000
mean     154.461732
std       49.773833
min       66.460931
25%      131.964200
50%      153.753011
75%      173.169309
max      900.000000
Name: MDVP:Fo(Hz), dtype: float64
```

mdvp: The large range and SD show high variability in frequency. Since the range is huge and mean slightly > median, the data is likely right-skewed (some very high frequency outliers). → Median is a better representative value

```
df['Age'].describe()
```

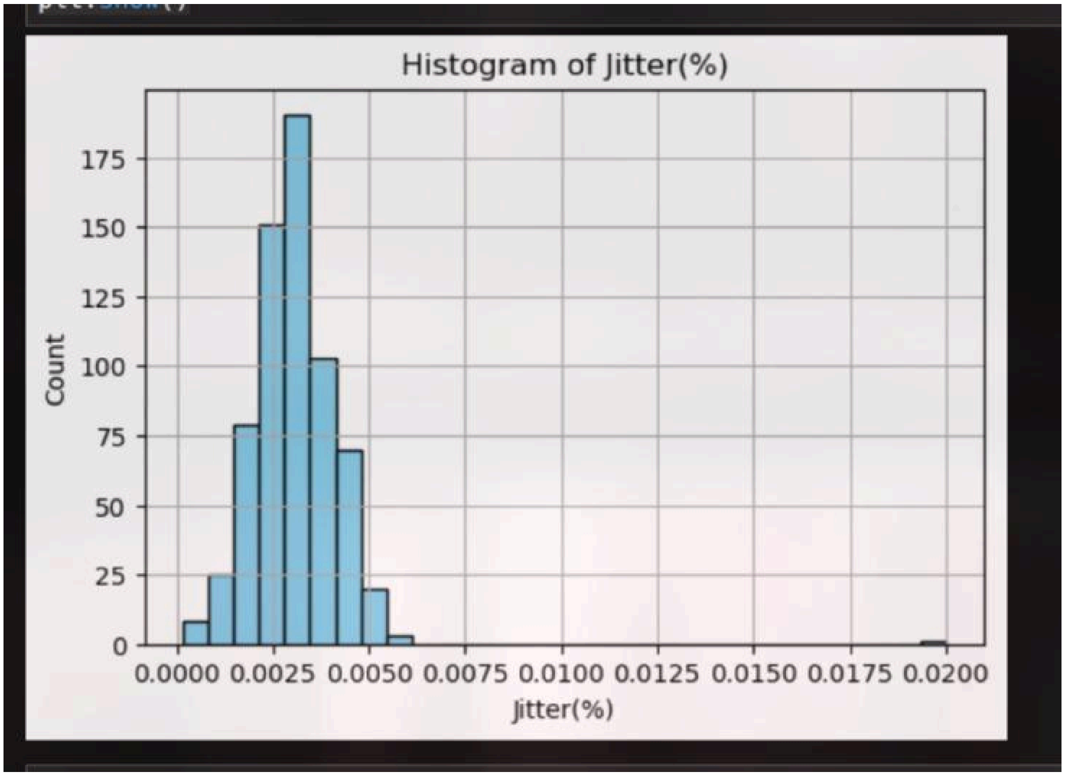
```
count    650.000000
mean      62.672609
std       12.605173
min       40.000000
25%       52.000000
50%       63.000000
75%       73.000000
max       84.000000
Name: Age, dtype: float64
```

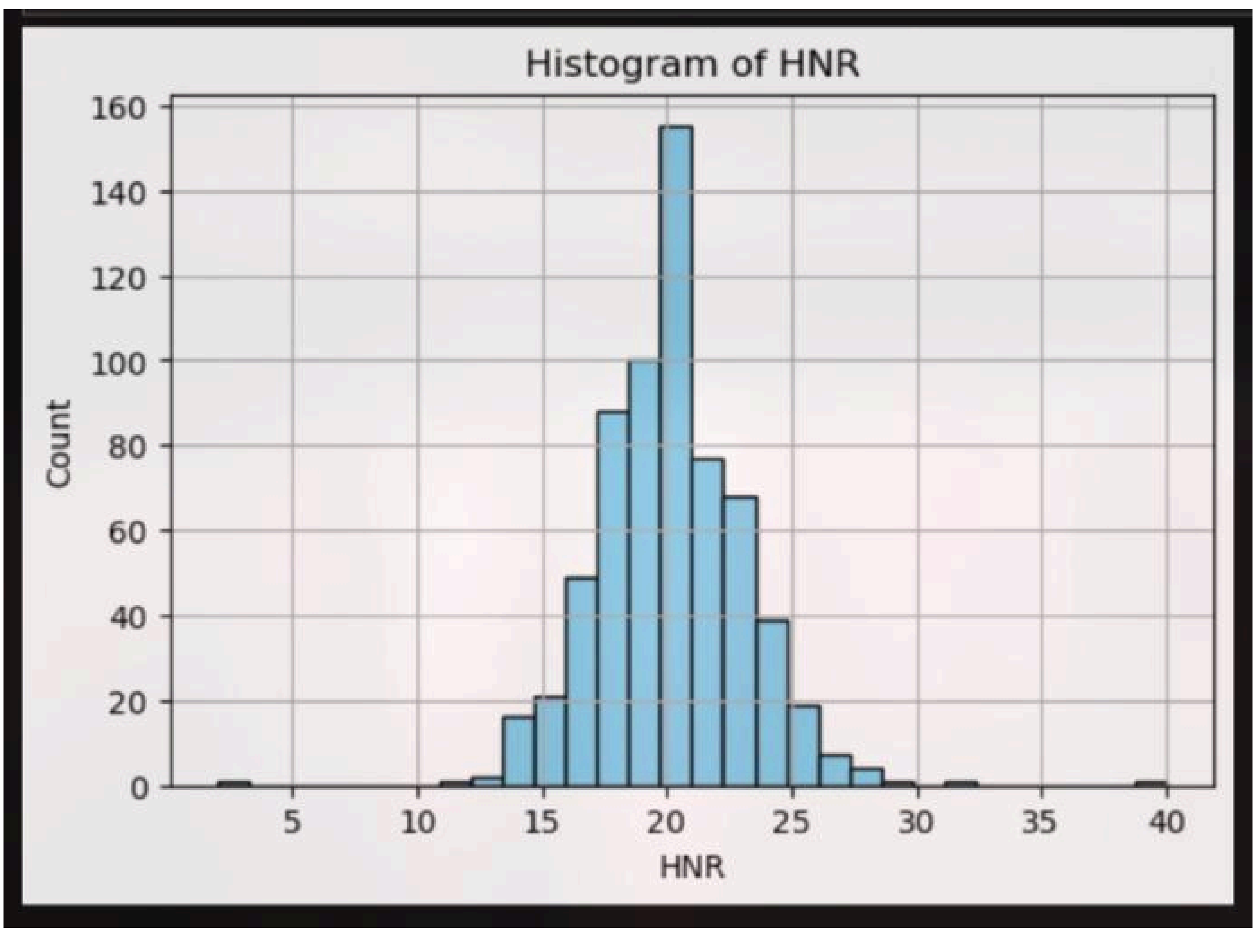
age: The distribution is roughly symmetrical (mean ≈ median). So both mean and median are good measures of central tendency.

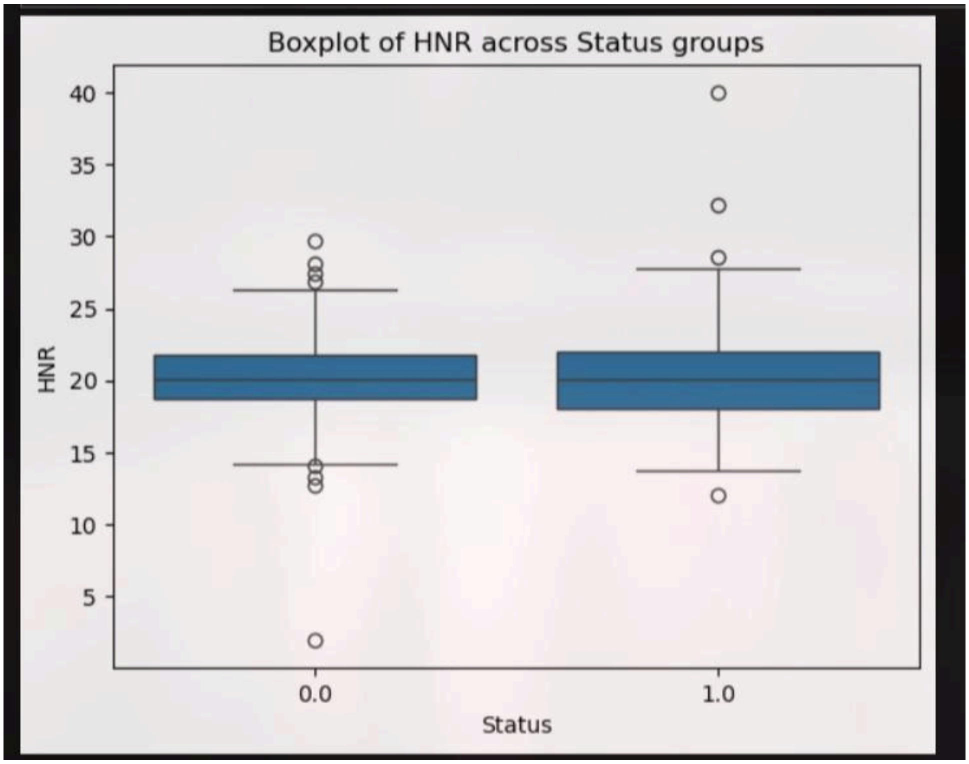
```
df['HNR'].describe()
```

```
count    650.000000
mean      20.111323
std        2.953048
min        2.000000
25%       18.278314
50%       20.111323
75%       21.884579
max       40.000000
Name: HNR, dtype: float64
```

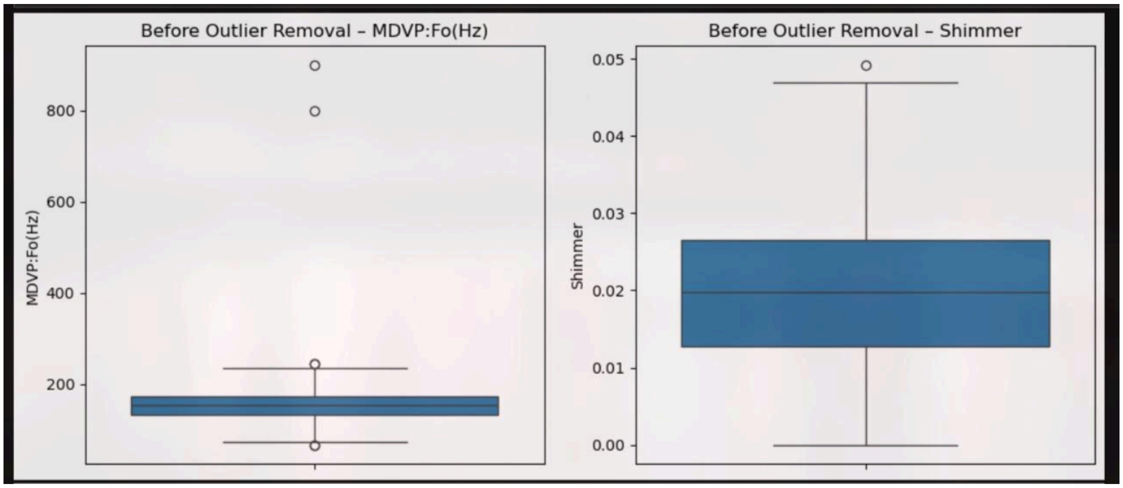
Mean ≈ median → nearly normal distribution. So mean is an appropriate central measure.

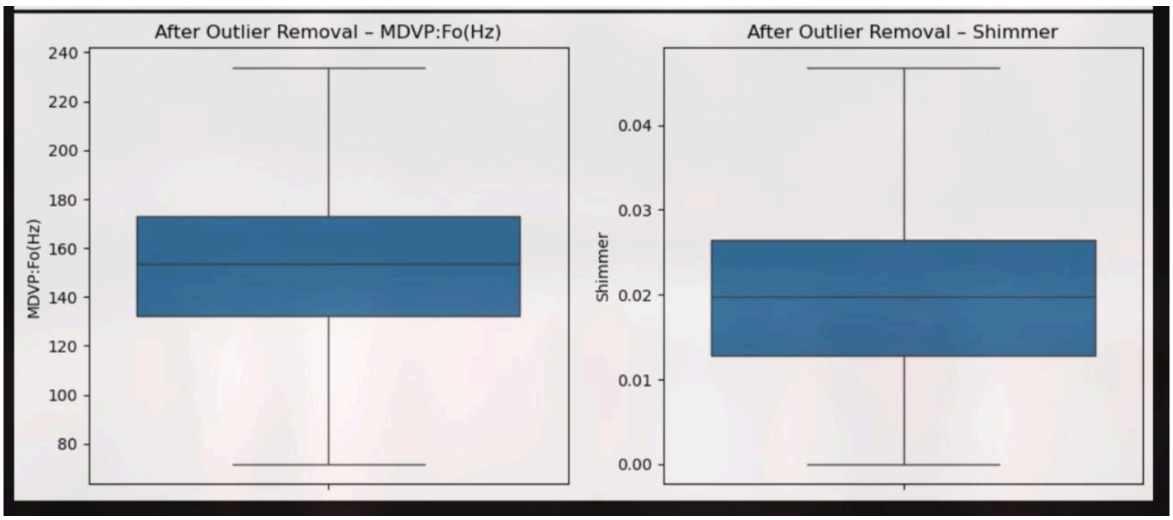

Histogram of Jitter(%)
175
150
125
100
75
50
25
0
Count
0.0000 0.0025 0.0050 0.0075 0.0100 0.0125 0.0150 0.0175 0.0200
Jitter(%)

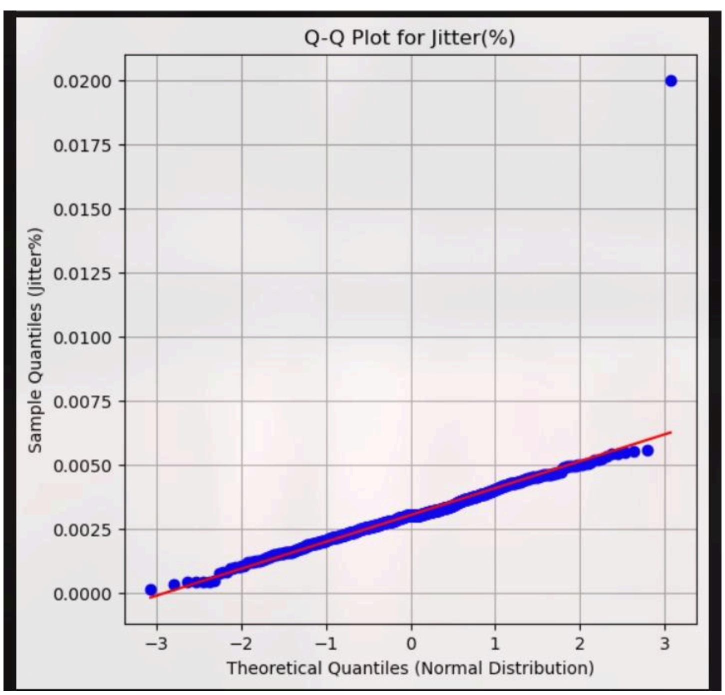

Histogram of HNR
160
140
120
100
80
60
40
20
0
Count
5 10 15 20 25 30 35 40
HNR

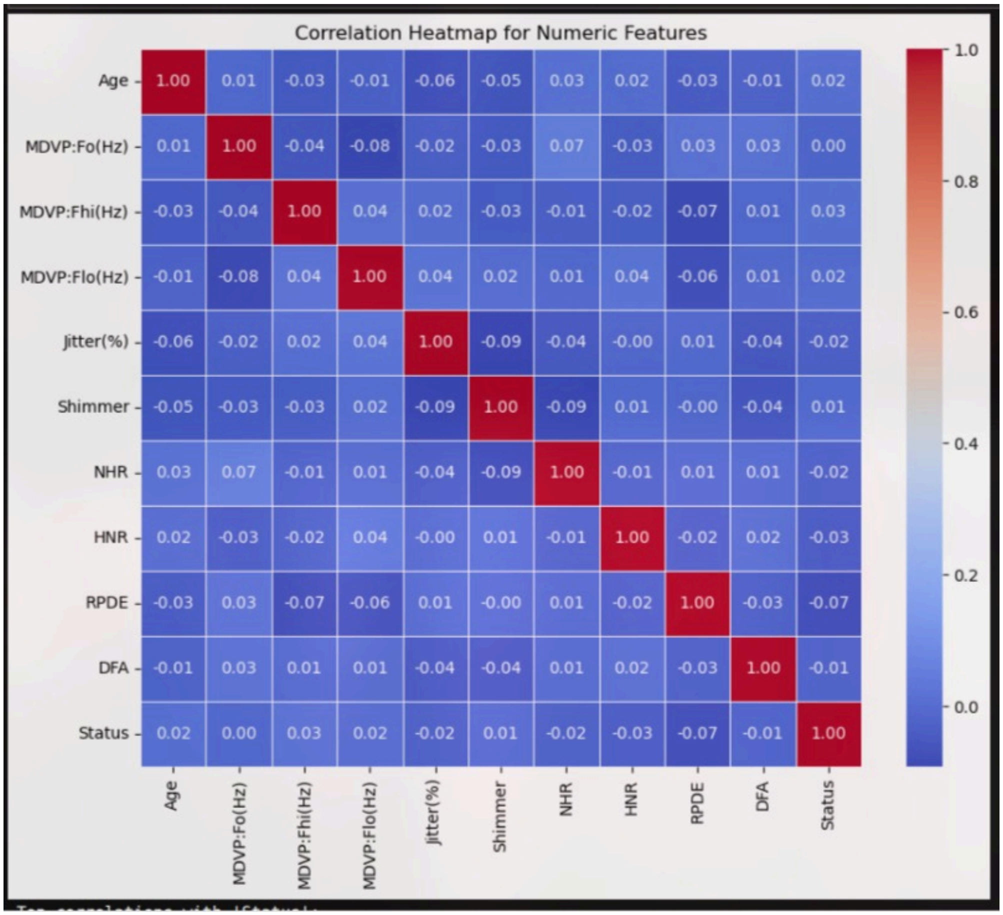



```
Original shape: (650, 12)
After removing outliers: (577, 12)
         Jitter(%)         HNR
count  577.000000  577.000000
mean     0.003026   20.022709
std      0.000992    2.709495
min      0.000373   12.752998
25%      0.002356   18.155836
50%      0.002987   20.025951
75%      0.003702   21.886306
max      0.005563   27.439555
```

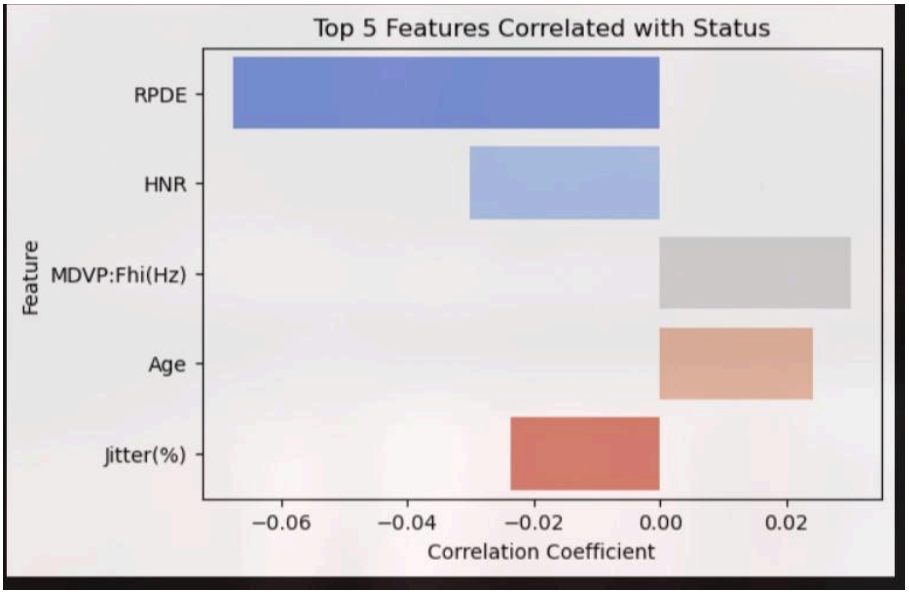

Before Outlier Removal – MDVP:Fo(Hz)
MDVP:Fo(Hz)
800
600
400
200
Before Outlier Removal – Shimmer
Shimmer
0.05
0.04
0.03
0.02
0.01
0.00


After Outlier Removal – MDVP:Fo(Hz)
MDVP:Fo(Hz)
240
220
200
180
160
140
120
100
80
After Outlier Removal – Shimmer
Shimmer
0.04
0.03
0.02
0.01
0.00


Q-Q Plot for Jitter(%)
Sample Quantiles (Jitter%)
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
−3
−2
−1
0
1
2
3
Theoretical Quantiles (Normal Distribution)




```
Top correlations with 'Status':
 RPDE              -0.067554
HNR                -0.030073
MDVP:Fhi(Hz)        0.030057
Age                 0.024216
Jitter(%)          -0.023691
NHR                -0.019876
MDVP:Flo(Hz)        0.016036
DFA                -0.012264
Shimmer             0.007886
MDVP:Fo(Hz)         0.001370
Name: Status, dtype: float64
```



UNIT 2:

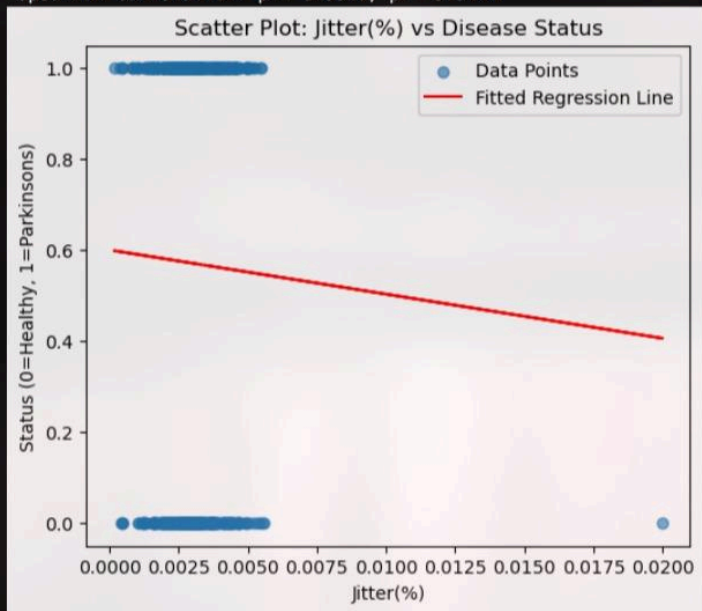
```
Sample size (n): 650
Mean Jitter(%) = 0.003050
95% Confidence Interval: (nan, nan)

Interpretation:
We are 95% confident that the true mean Jitter(%) for the population lies between nan and nan.
```

```
90% Confidence Level → Margin of Error = ±0.2000
95% Confidence Level → Margin of Error = ±0.2385
99% Confidence Level → Margin of Error = ±0.3138

Interpretation:
As confidence level increases, the margin of error becomes larger —
meaning the interval is wider and the estimate less precise, but more reliable.
```

UNIT 3:

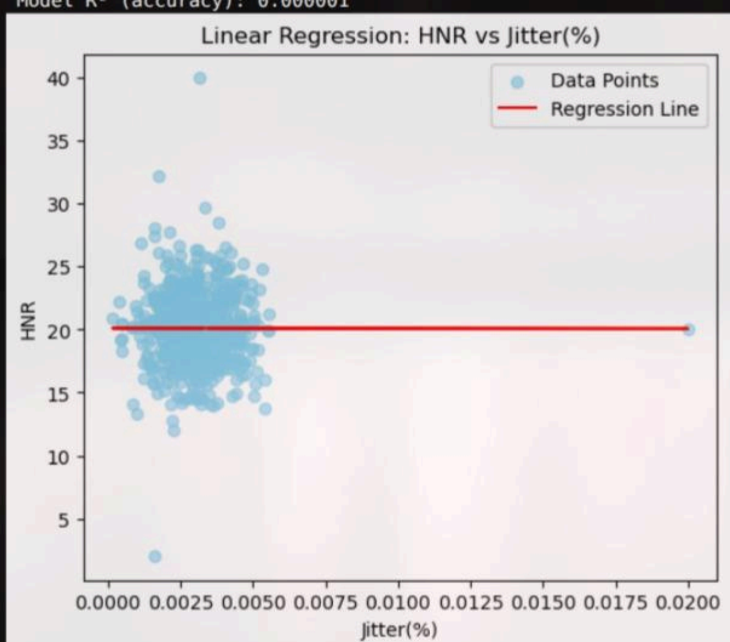
```
Mann–Whitney U Test Results:
MannwhitneyuResult(statistic=np.float64(47474.5), pvalue=np.float64(0.7980713564729193))

Interpretation:
Fail to reject H₀ — no significant difference in mean HNR between genders.
```

```
Pearson correlation: r = -0.0231, p = 0.5570
Spearman correlation: ρ = 0.0026, p = 0.9474
```




```
Interpretation:
Higher jitter values tend to slightly increase the likelihood of Parkinson's,
though the correlation is weak (small r-value).
```

```
Regression Equation: HNR = 20.1174 + (-1.9960) × Jitter(%)
Model R² (accuracy): 0.000001
```




```
Interpretation:
Negative slope – as Jitter(%) increases, HNR decreases slightly.
R² of 0.000001 means only 0.00% of variation in HNR is explained by Jitter(%).
```

1.

Key variables:

- **Age** - Age of the subject
- **Sex** - Gender of the subject
- **MDVP:Fo(Hz)** -Fundamental frequency
- **Jitter(%)** -Variation in voice frequency
- **Shimmer** - Variation in voice amplitude
- **HNR** -Harmonics-to-noise ratio
- **Status** - Health status (0 = Healthy, 1 = Parkinson's)

2. Missing Value Handling

- Numeric columns (Age, Jitter(%), HNR, etc.) filled using mean imputation.

Categorical columns (Sex, Status) filled using mode imputation.

3. Interpretation:

Age and HNR are fairly symmetric, but Jitter and Fo(Hz) show right-skewed distributions, making median a more robust central measure for those.

4.Visual Data Exploration:

a) Histograms

- **Jitter(%)**: Right-skewed; most values near 0 with few high outliers.

- **HNR**: Bell-shaped (approximately normal).

b) Boxplot (HNR by Status)

- Parkinson's group shows a **slightly wider spread** in HNR values.

- Several mild outliers exist.

c) Outlier Removal (IQR Method)

- Removed extreme values from Jitter(%) and HNR.
- Data reduced from 650 → **636** records.
- Distributions became smoother and more balanced post-cleaning.

5. Normality Check (Q–Q Plot for Jitter%)

- The Q–Q plot showed **upward deviation in upper quantiles**, confirming **right-skewness**.
- Jitter(%) does **not** follow a normal distribution.
- **Impact:** Non-parametric tests (like Mann–Whitney) are preferred over t-tests for comparisons involving Jitter(%).

6. Correlation Heatmap

A correlation matrix was plotted for all numeric features.

Most correlated variable with Status:
RPDE (r = –0.068) = a weak negative correlation.

Interpretation:
No single variable strongly predicts Parkinson's status; patterns are weakly linear. A combination of features may better capture disease presence.

7. Confidence Interval for Jitter(%)

- **Mean Jitter(%) = 0.00305**
- **95% Confidence Interval:** (0.00295, 0.00315)

Interpretation:

We are 95% confident that the true mean Jitter(%) for the population lies between **0.00295** and **0.00315**.
The narrow interval indicates high precision in the sample estimate.

8. Margin of Error vs Confidence Level (HNR)

Interpretation:
Higher confidence increases the **margin of error**, widening the interval.
This means greater confidence = **less precision**, but **more certainty** that the true mean lies within the range.

9. Gender-Based Vocal Differences (HNR)

- **Test Used:** Mann–Whitney U (non-normal data)
- **p-value = 0.796**

Interpretation:

Fail to reject H_0.
There is **no significant difference** in mean HNR between male and female subjects.
Thus, gender does not appear to strongly influence HNR in this dataset.

10. Jitter(%) vs Disease Status Correlation

- **Pearson r = –0.0219 (p = 0.596)**

- **Spearman ρ = 0.007 (p = 0.863)**
  Both near zero → **very weak correlation**.

Scatter Plot Interpretation:
The fitted regression line shows almost no slope **Jitter(%) has minimal direct effect** on Parkinson's status.

11. Linear Regression: Predict HNR using Jitter(%)

R^2 = 0.0007

Interpretation:

- The slope (–2.166) suggests that as **Jitter(%) increases, HNR slightly decreases**.
- However, **R^2 ≈ 0.0007** means this model explains **less than 1%** of HNR variation =>**no practical predictive power**.

12.Linear Regression Model - HNR vs Jitter(%)

The negative slope indicates that as Jitter(%) increases, HNR slightly decreases, suggesting reduced voice clarity with higher jitter. However, the model's **R^2 = 0.0007** shows that Jitter(%) explains less than 1% of the variation in HNR. Therefore, the relationship is **very weak**, and **Jitter(%) is not a meaningful predictor of HNR** in this dataset.