Case Study - Loan Data Analysis

This dataset contain complete loan data for all loans issued through the 2018 Q4, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing data with additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections among others.

Problem Statement 1

1. Print all the columns names in a loop
2. Create a DataFrame with the columns specified
"term", "home_ownership", "grade", "purpose", "int_rate",
"installment","addr_state","loan_status","application_type","loan_amnt","emp_length",
"annual_inc","dti","delinq_2yrs","revol_bal","revol_util","total_acc","num_tl_90g_dpd_24m","dti_joint"

```

```

Problem Statement 2

Compute basic statistics for the column 'loan_amnt' and 'annual_inc'

Expected Output

```
+-------+------------------+------------------+
|summary|         loan_amnt|        annual_inc|
+-------+------------------+------------------+
|  count|            128412|            128412|
|   mean| 15971.32102139987|  82797.3278609475|
| stddev|10150.384232741928|108298.46579150086|
|    min|              1000|               0.0|
|    max|             40000|         9757200.0|
+-------+------------------+------------------+
```

```

```

Problem Statement 3

Show distinct values of 'emp_length' column

Expected output

```
+----------+
|emp_length|
+----------+
|   5 years|
|   9 years|
|      null|
|    1 year|
|       n/a|
|   2 years|
|   7 years|
|   8 years|
|   4 years|
|   6 years|
|   3 years|
| 10+ years|
|  < 1 year|
+----------+
```

Problem Statement 4

If you observe the column emp_length the data contains not just numeric but also string like 'years' and also special characters. Create a column by name 'emplength_cleaned' by having only integers

```
+-----------------+----------+
|emplength_cleaned|emp_length|
+-----------------+----------+
|                1|  < 1 year|
|               10| 10+ years|
|                1|  < 1 year|
|              n/a|       n/a|
|                5|   5 years|
|                9|   9 years|
|                3|   3 years|
|               10| 10+ years|
|              n/a|       n/a|
|               10| 10+ years|
+-----------------+----------+
only showing top 10 rows
```