Career Foundry
Data Analytics Portfolio

Achievement 6: Advanced Analytics and Dashboard Design

Julie Spigner

June 2021

kiva

Achievement 6 Project

Kiva is an online crowdfunding platform to extend financial services to underserved and financially excluded people around the world.

Kiva lenders have provided over **$1 billion dollars** in loans to over **2 million people**.

In order to set investment priorities, help inform lenders, and understand their target communities, knowing the level of poverty of each borrower is critical.

Questions

How does one qualify for a loan with Kiva?
Who are currently getting loans with Kiva?
What are the loans used for?
Where are loans distributed to and how are they determined?
What is the average in loan amount given to borrowers?

Data Source

The source of the data is internal and Kiva owns it. There are four data sets to work with and they consist information about loans, locations, loan themes, and loan themes by region.

Data sets:

⛓️ **kiva_loans.csv**
⛓️ **kiva_mpi_region_locations.csv**
⛓️ **loan_theme_ids.csb**
⛓️ **loan_themes_by_region.csv**

kiva_loans.csv

- 👉Is the most valuable and comprehensive data set

- 👉Contains details about activities the funds are used for, where and whom they're distributed to, how much is being borrowed and repaid, and the length of terms of loans

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 671205 entries, 0 to 671204
Data columns (total 20 columns):
 #   Column              Non-Null Count   Dtype  
---  ------              --------------   -----  
 0   id                  671205 non-null  int64  
 1   funded_amount       671205 non-null  float64
 2   loan_amount         671205 non-null  float64
 3   activity            671205 non-null  object 
 4   sector              671205 non-null  object 
 5   use                 666973 non-null  object 
 6   country_code        671197 non-null  object 
 7   country             671205 non-null  object 
 8   region              614405 non-null  object 
 9   currency            671205 non-null  object 
 10  partner_id          657698 non-null  float64
 11  posted_time         671205 non-null  object 
 12  disbursed_time      668809 non-null  object 
 13  funded_time         622874 non-null  object 
 14  term_in_months      671205 non-null  float64
 15  lender_count        671205 non-null  int64  
 16  tags                499789 non-null  object 
 17  borrower_genders    666984 non-null  object 
 18  repayment_interval  671205 non-null  object 
 19  date                671205 non-null  object 
dtypes: float64(4), int64(2), object(14)
memory usage: 102.4+ MB
```

kiva_mpi_region_locations.csv

👉Provides geospatial information on loans

👉Uses **International Organization for Standardization (ISO)** as a system to identify locations of loans

👉Is used to create a map in Tableau

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2772 entries, 0 to 2771
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   LocationName  984 non-null    object
 1   ISO           1008 non-null   object
 2   country       1008 non-null   object
 3   region        984 non-null    object
 4   world_region  1008 non-null   object
 5   MPI           984 non-null    float64
 6   geo           2772 non-null   object
 7   lat           892 non-null    float64
 8   lon           892 non-null    float64
dtypes: float64(3), object(6)
memory usage: 195.0+ KB
```

loan_theme_ids.csv

👉Is a subset of kiva_loans.csv

👉Contains records of activities the loans are used for

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 779092 entries, 0 to 779091
Data columns (total 4 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   id                 779092 non-null  int64
 1   Loan Theme ID      764279 non-null  object
 2   Loan Theme Type    764279 non-null  object
 3   Partner ID         764279 non-null  float64
dtypes: float64(1), int64(1), object(2)
memory usage: 23.8+ MB
```

loan_themes_by_region.csv

👉Links loans with locations

👉Uses **Multidimensional Poverty Index** (MPI) as a measurement tool for loan eligibility

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15736 entries, 0 to 15735
Data columns (total 21 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Partner ID          15736 non-null  int64
 1   Field Partner Name  15736 non-null  object
 2   sector              15736 non-null  object
 3   Loan Theme ID       15736 non-null  object
 4   Loan Theme Type     15736 non-null  object
 5   country             15736 non-null  object
 6   forkiva             15736 non-null  object
 7   region              15736 non-null  object
 8   geocode_old         1200 non-null   object
 9   ISO                 15722 non-null  object
 10  number              15736 non-null  int64
 11  amount              15736 non-null  int64
 12  LocationName        15736 non-null  object
 13  geocode             13662 non-null  object
 14  names               13661 non-null  object
 15  geo                 15736 non-null  object
 16  lat                 13662 non-null  float64
 17  lon                 13662 non-null  float64
 18  mpi_region          15722 non-null  object
 19  mpi_geo             9671 non-null   object
 20  rural_pct           14344 non-null  float64
dtypes: float64(3), int64(3), object(15)
memory usage: 2.5+ MB
```

Python Scripts via Jupyter Notebook

🐍 Kiva – Data Cleaning.ipynb
🐍 Kiva – Exploring Relationships.ipynb
🐍 Kiva – Exploratory Analysis.ipynb
🐍 Kiva – Geographic Visualization.ipynb
🐍 Kiva – Machine Learning – Regression Analysis.ipynb
🐍 Kiva – Machine Learning – Clustering.ipynb
🐍 Kiva – Time-Series Data.ipynb

Scripts are also uploaded in GitHub

Exploratory Analysis

- There are **163 activities** listed.

- The **top 5** most popular activities for loans are as followed: farming, general store, personal housing expenses, food production/sales, and agriculture.

- The **bottom 5** least popular activities for loans are as followed: film, personal care products, event planning, celebrations, and adult care.

	activity
Farming	72955
General Store	64729
Personal Housing Expenses	32448
Food Production/Sales	28106
Agriculture	27023
...	...
Film	13
Personal Care Products	7
Celebrations	5
Event Planning	5
Adult Care	2

Exploratory Analysis

Notes: According to this list, the **top 10 countries** with the most loan count in place are as followed:

Philippines
Armenia
Colombia
Mexico
Vietnam
Peru
Kenya
Cambodia
El Salvador
Tajikistan

	ISO
PHL	3467
ARM	1064
COL	824
MEX	716
VNM	687
PER	609
KEN	593
KHM	567
SLV	496
TJK	454

Exploratory Analysis

I was puzzled to see **"Farming"** and **"Agriculture"** listed as separate activities so, I conducted some quick research to see if there's a difference between the two. According to Kiva's official website, farming is a subcategory under Agriculture.

I looked up current loans in **Philippines**. As of May 2021, there are **1,574 active loans**. The majority of those loans are related to **farming, agriculture, and pigs**.

	activity
Farming	72955
General Store	64729
Personal Housing Expenses	32448
Food Production/Sales	28106
Agriculture	27023
Pigs	26624
Retail	24771
Clothing Sales	22339
Home Appliances	20267
Higher education costs	19742

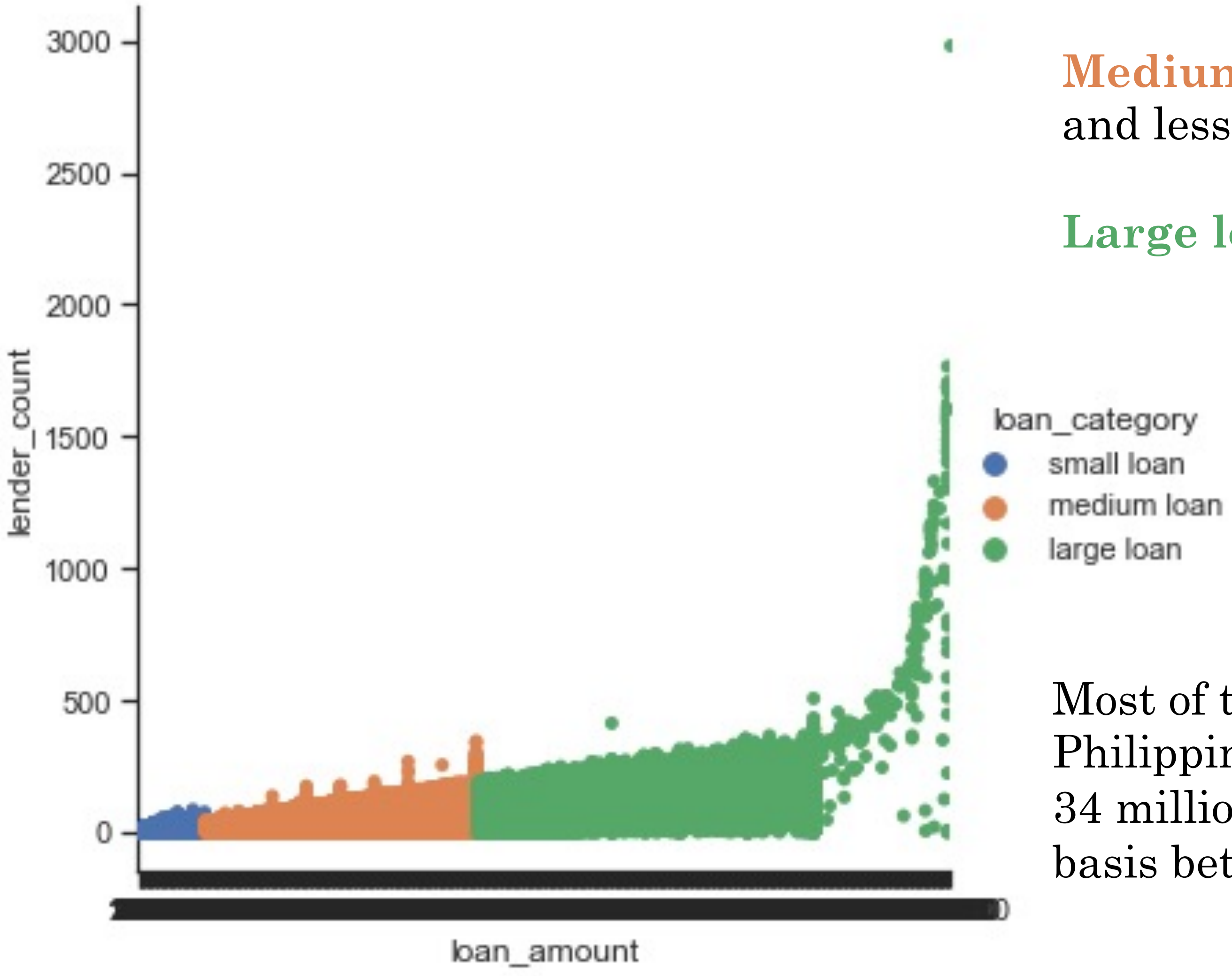
Loans Insights

Categories are created based on “**loan_amount**”: Small loan, medium loan, and large loan.

Small loans are less than or equal to $1,000.

Medium loans are greater than or equal to $1,001 and less than $5,001.

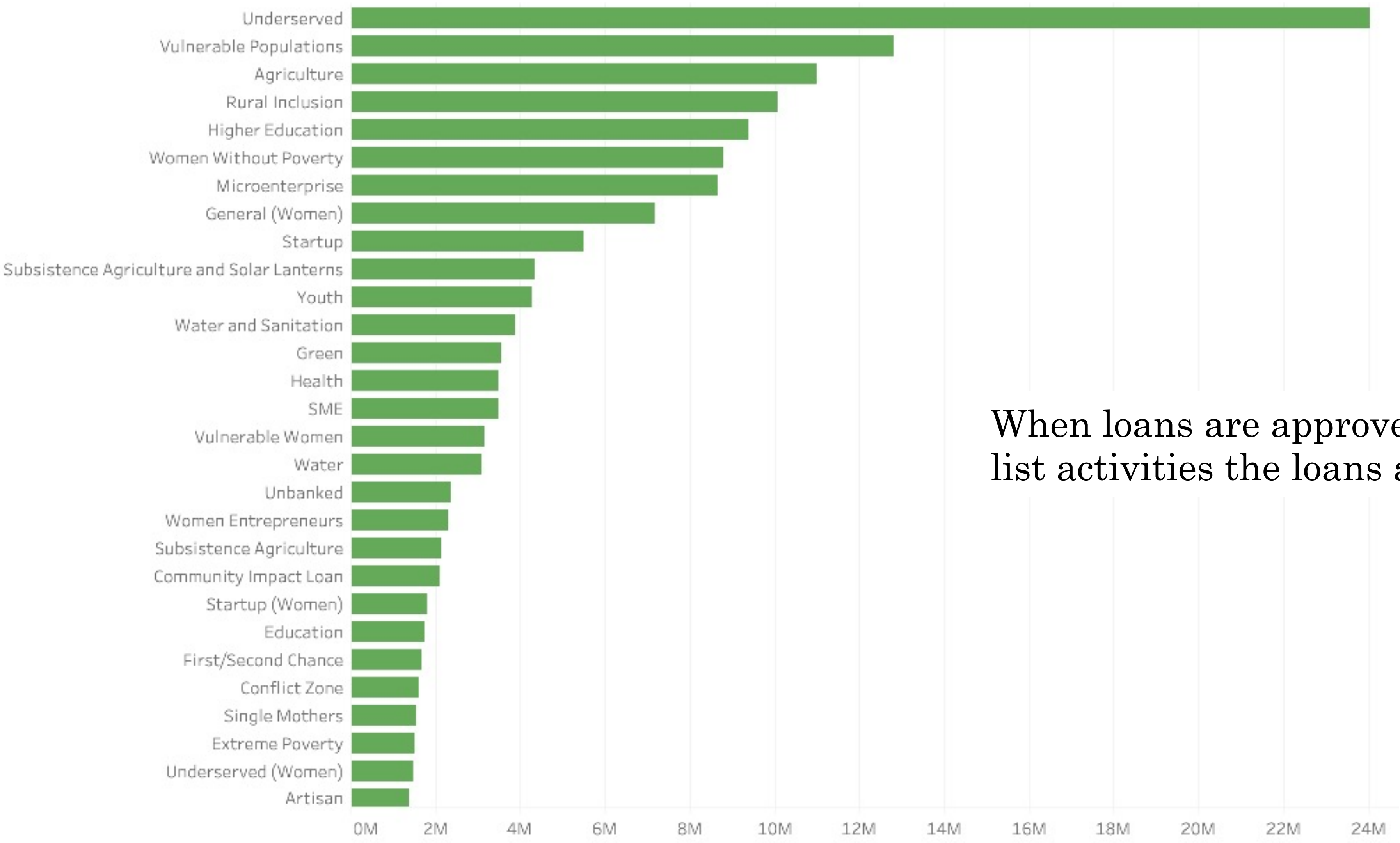
Large loans are greater than or equal to $5,001.




```
small loan      521512
medium loan     140114
large loan        9579
Name: loan_category, dtype: int64
```

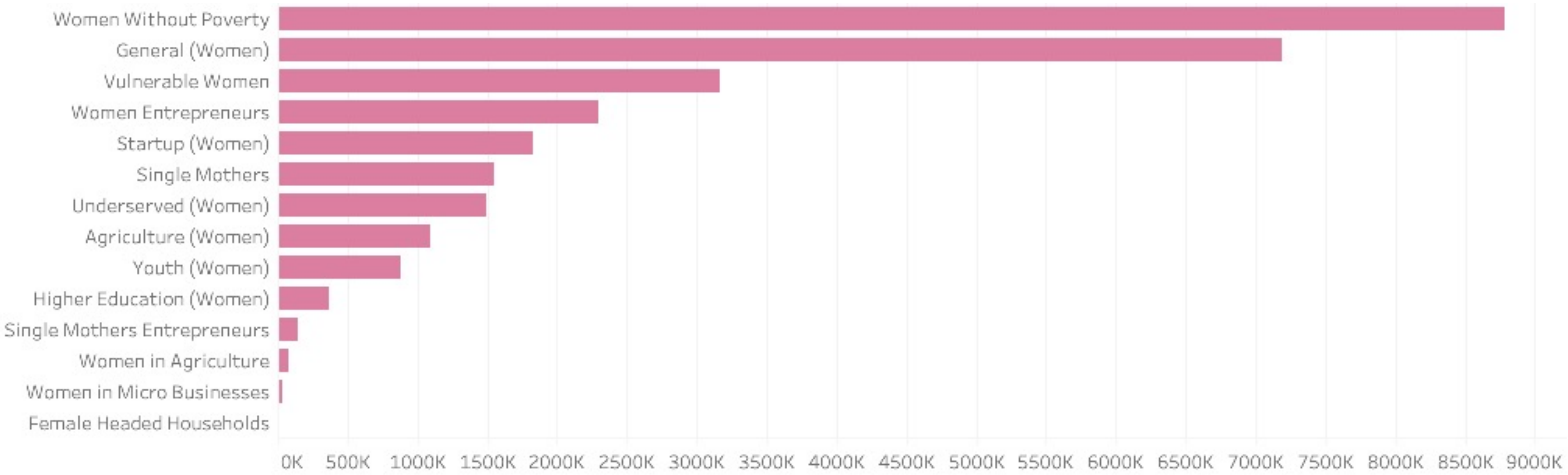
Most of the loans are distributed primarily in Philippines with just a little over 3,000 loans totaling 34 million USD. Terms of loans are on a case-by-case basis between 0 to 72 months.

Loans Insights



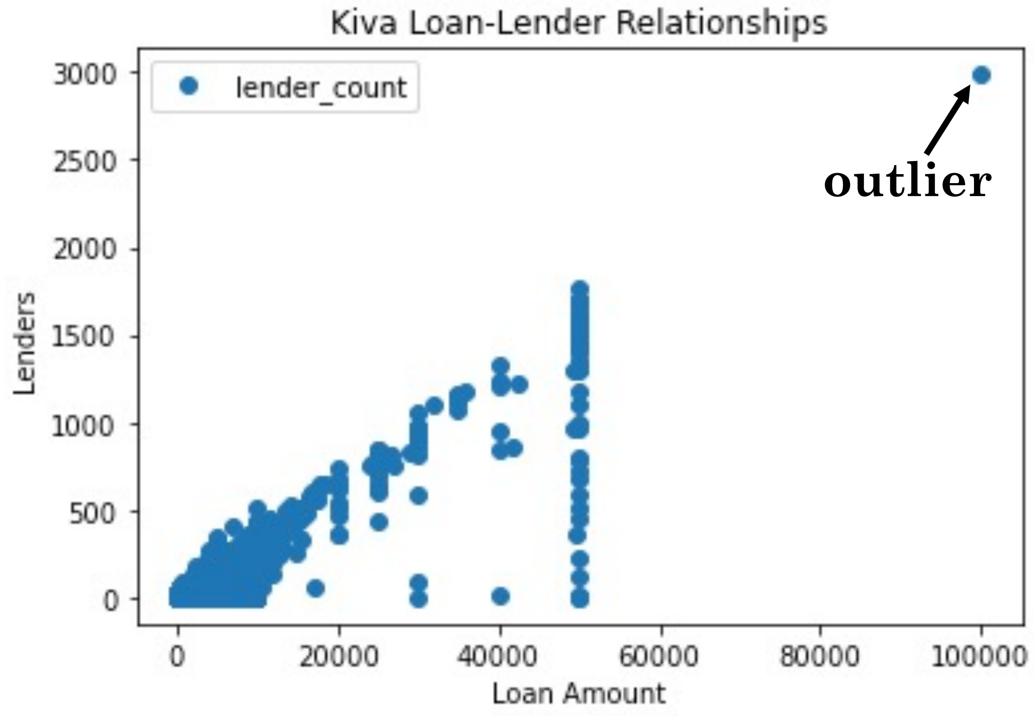
When loans are approved, borrowers list activities the loans are used for.

Loans Insights

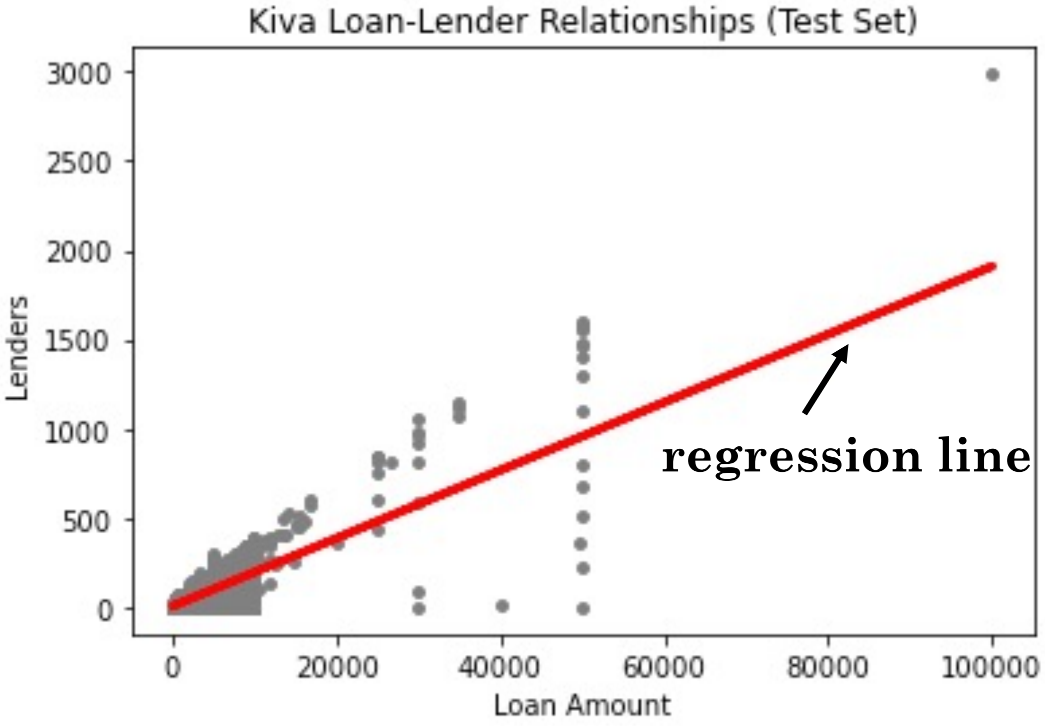


Some loan activities are female-specific or borrowed by female borrowers.

Regression Analysis

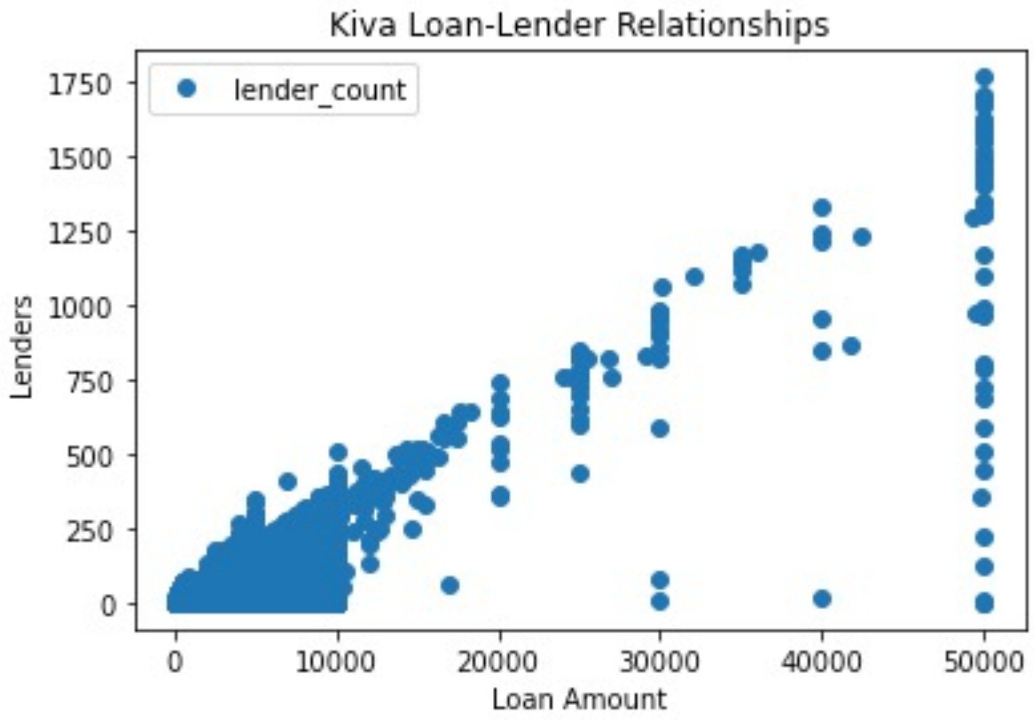


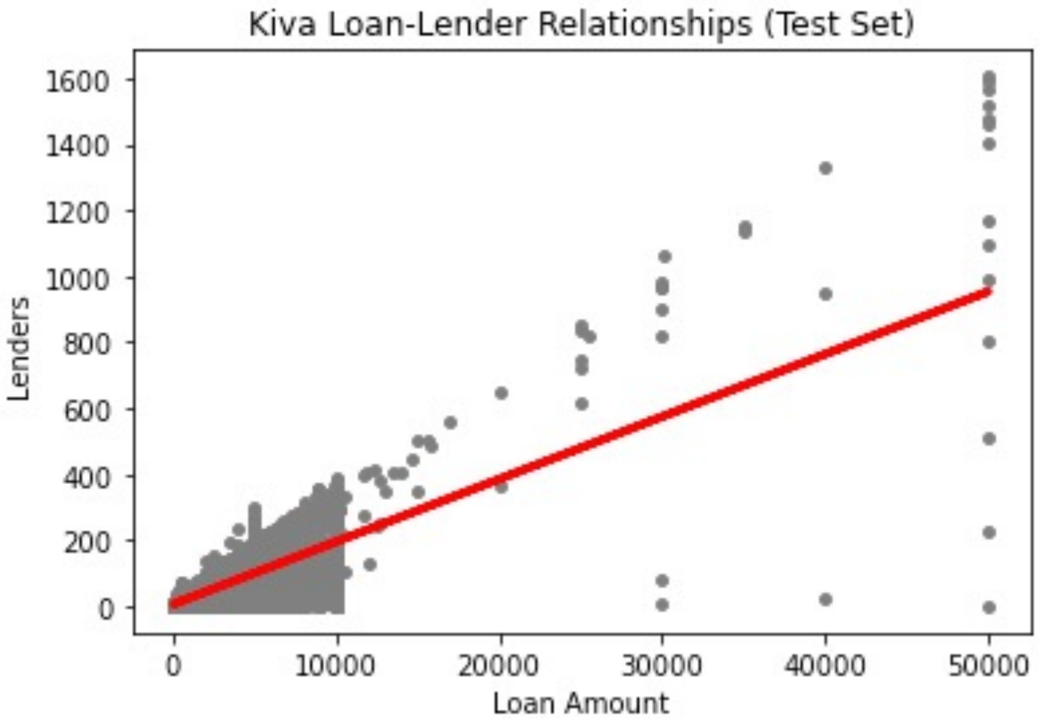
There was one outlier: a group of 3,000 borrowers joined together for a loan of $100.000.



The outlier skewed the results. There were also multiple data points outside of the regression line.

Regression Analysis

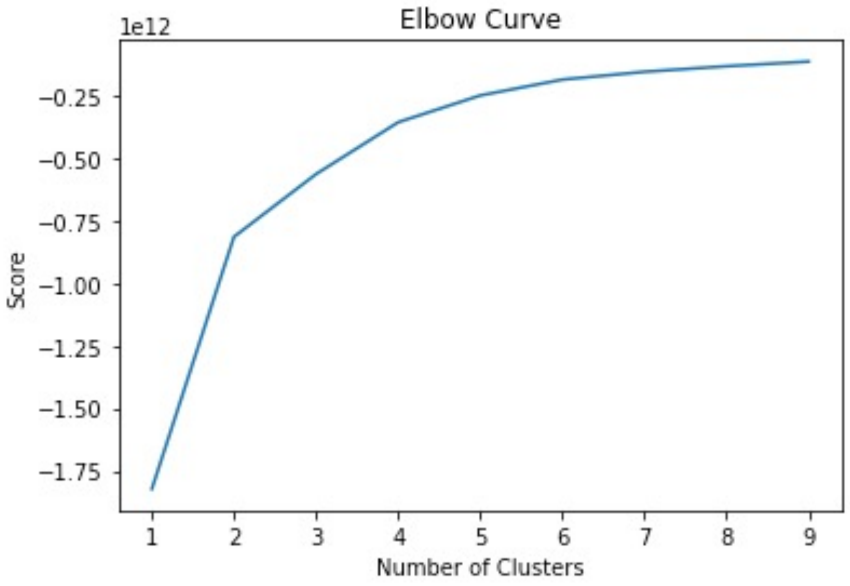


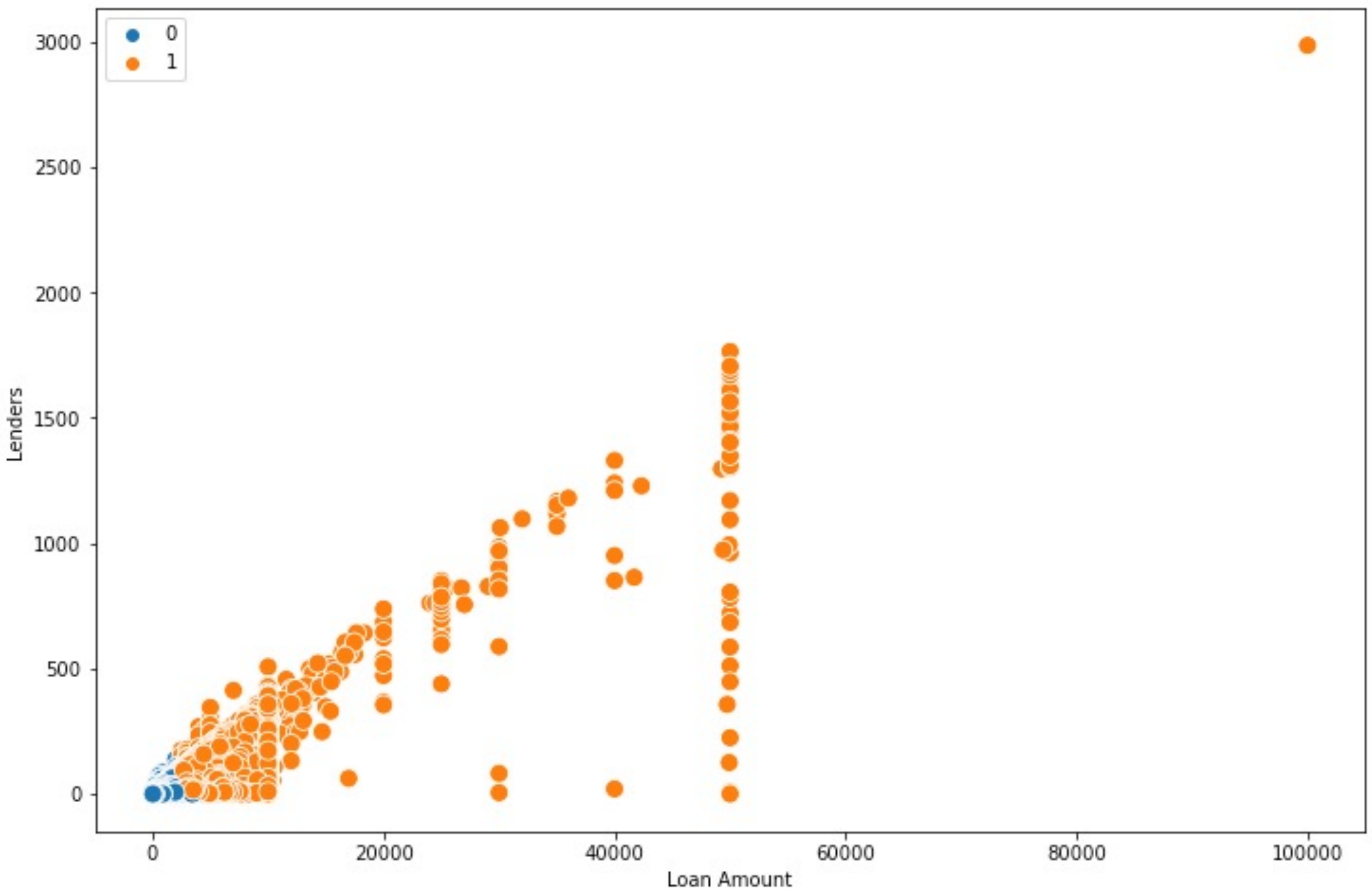


Then, the outlier was removed and the linear regression was tested again. Even though the outlier was removed, the results did not improve.

There were still multiple data points outside of the regression line. The R2 Score for both models were weak (at 0.6). This suggests that a linear regression is not an appropriate model to predict Loan-Lender relationships.

Cluster Analysis



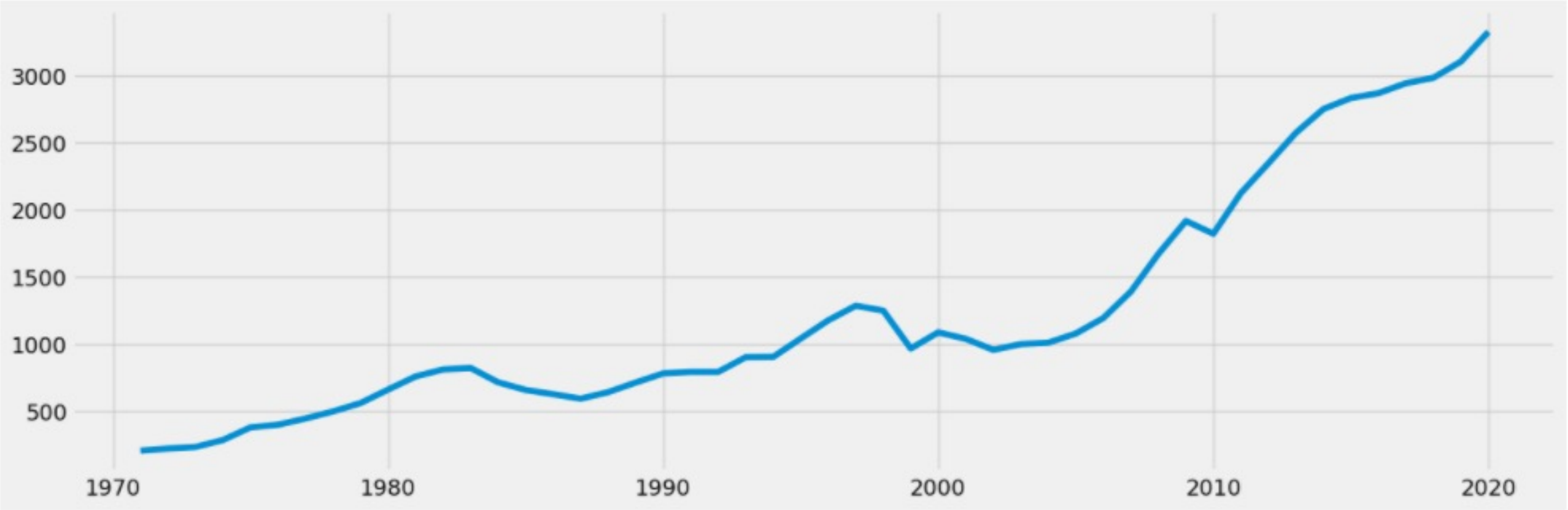


In order to prepare for the k-clustering algorithm, the **Elbow Curve technique** (left) was applied to check how many clusters there are. A large jump from 1 to 2 on the x-axis is visible, but after that, the curve straightens out. This means the optimal count for the clusters is 2.

The **k-clustering chart** (right) shows data points clustered in two clusters. **Blue clusters** indicate low number of lenders and low number of loans. **Orange clusters** represent larger number of lenders grouped together for bigger loans.

Time-Series

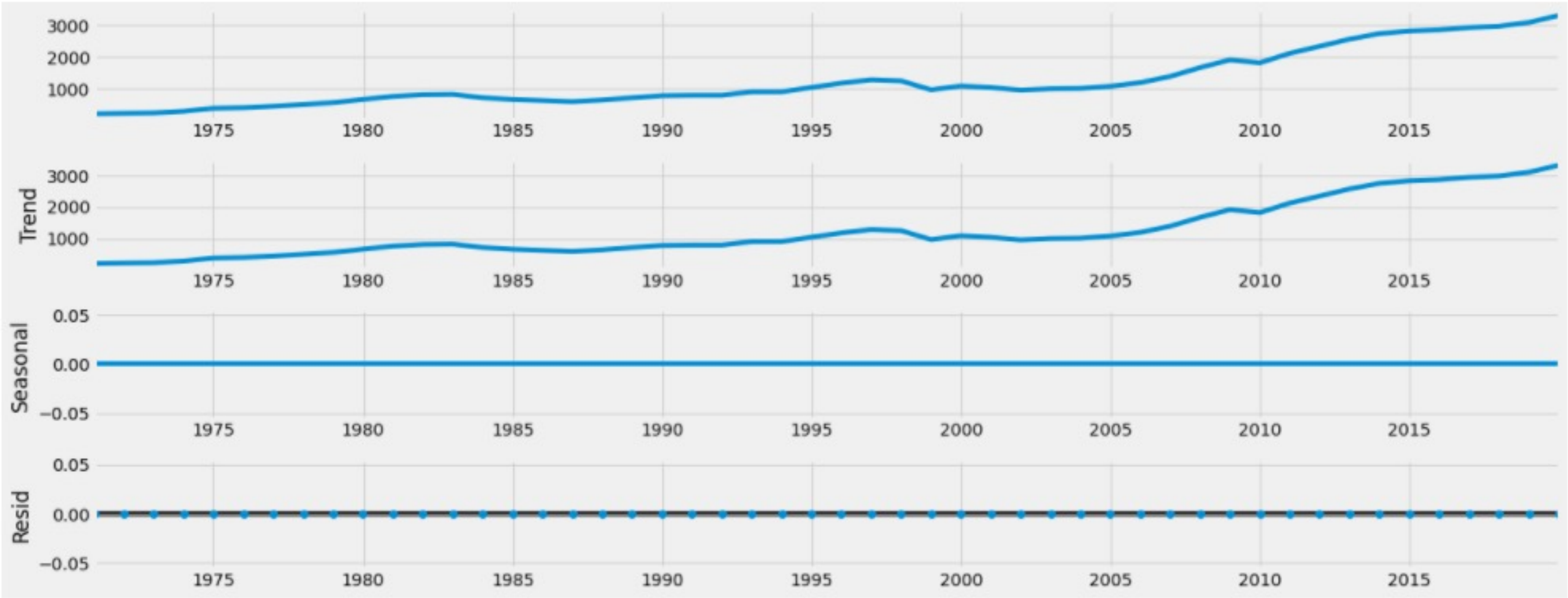
All time series charts were created in Python via Jupyter Notebook.



This data source was archived from **Quandl**. Philippines is the leading country with the most loans in place with Kiva. Farming and agriculture are their top two activities. An overview of time series data exists: Gross Domestic Product and year. GDP gives information about Philippines' economy and how they're faring compared to other countries. There is an **upward trend** from 1970 to 2020.

Data: United Nations Commodity Trade: Trade in Edible Vegetables and Certain Roots and Tuber – Philippines

Time-Series

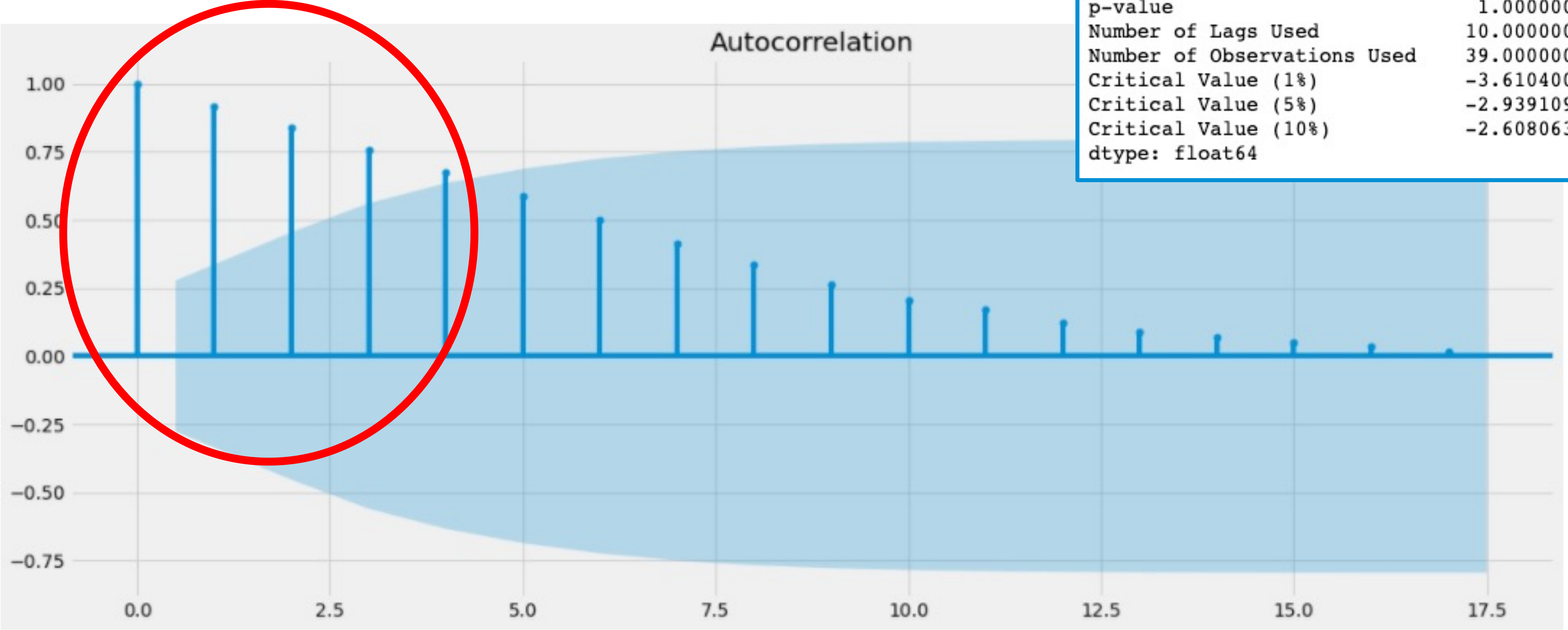


The **decomposition** (additive model) was applied. The trend shows an upward trend. There's no seasonal data. This suggests that the economy has been improving over time in Philippines, though slowly.

Data: United Nations Commodity Trade: Trade in Edible Vegetables and Certain Roots and Tuber – Philippines

Time-Series

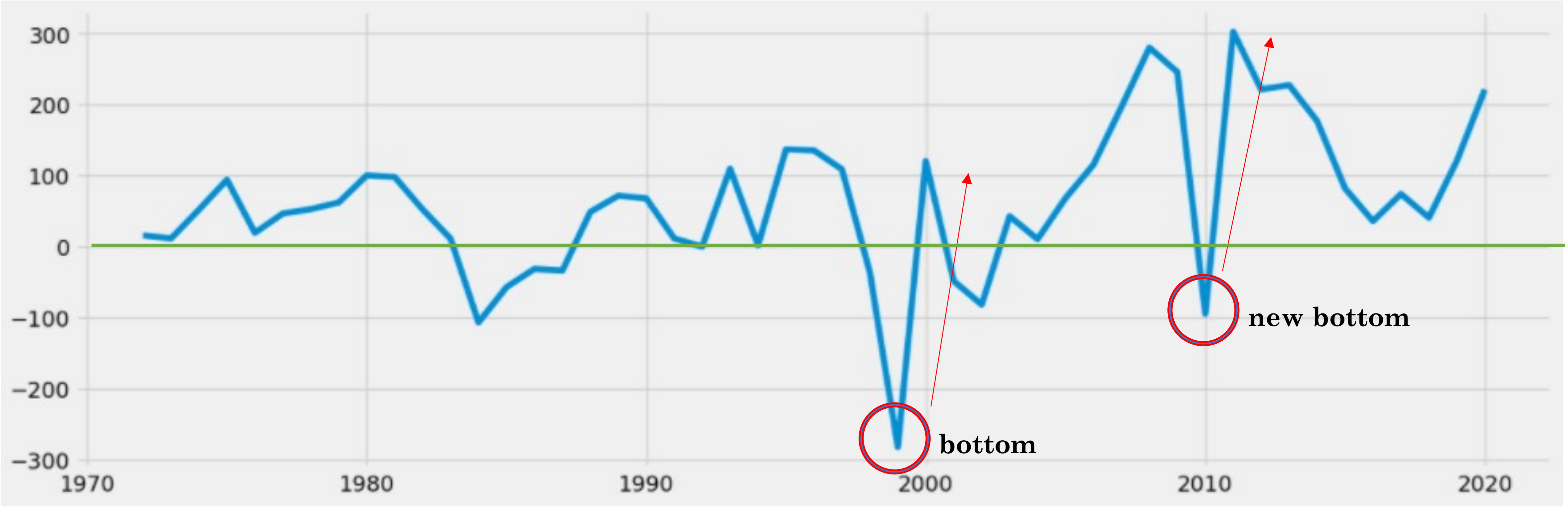
```
Dickey-Fuller Stationarity test:
Test Statistic                  3.408224
p-value                         1.000000
Number of Lags Used            10.000000
Number of Observations Used    39.000000
Critical Value (1%)            -3.610400
Critical Value (5%)            -2.939109
Critical Value (10%)           -2.608063
dtype: float64
```



In order to test for stationarity, the **Dickey-Fuller Stationary test** was applied. The vertical lines represent the confidence interval. There are multiple lines that go above the blue edge of the confidence interval which means there are some lags that are significantly correlated with each other.

Data: United Nations Commodity Trade: Trade in Edible Vegetables and Certain Roots and Tuber - Philippines

Time-Series

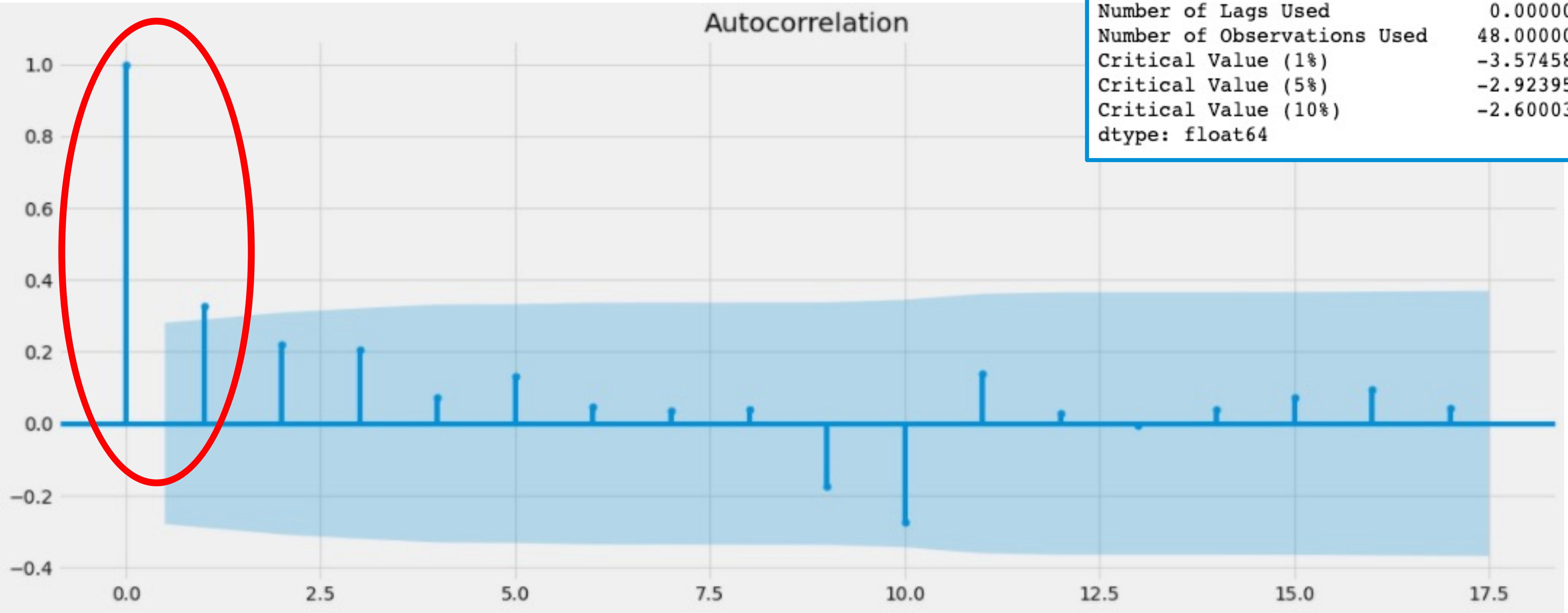


Here's an overview of what the differencing did to the time-series curve. There were two sharp dips in 1999 and 2010 in Philippines' economy. The dips indicate events of economic crash. Regardless of the dips, the economy recovered quickly thereafter. Currently, data is being collected to configure how the pandemic (COVID-19) impacts the local and global economy and has not been updated since January 2020.

Data: United Nations Commodity Trade: Trade in Edible Vegetables and Certain Roots and Tuber - Philippines

Time-Series

```
Dickey-Fuller Stationarity test:
Test Statistic                  -4.619371
p-value                          0.000119
Number of Lags Used              0.000000
Number of Observations Used     48.000000
Critical Value (1%)             -3.574589
Critical Value (5%)             -2.923954
Critical Value (10%)            -2.600039
dtype: float64
```



In order to test for stationarity after differencing, the **Dickey-Fuller Stationary test** was applied again. This time, there are only two lines that go above the blue edge of the confidence interval which means there are a few lags that are significantly correlated with each other.

Data: United Nations Commodity Trade: Trade in Edible Vegetables and Certain Roots and Tuber - Philippines

Recommendations

- 👉 Categorize types of loans to gain more insights
- 👉 Organize activities into fixed categories to find more trends
- 👉 Gather more information about borrowers
- 👉 Work as a team with…
  - 👉 financial analysts (or economists) to interpret findings in global economics
  - 👉 humanitarians to identify problems and solutions for loans with multiple borrowers
- 👉 Run more tests to find additional factors that impact Loan-Lender relationships

Links

🔗 GitHub
🔗 Kiva
🔗 Kaggle
🔗 Tableau
🔗 Qandl

Contact me if you have questions, suggestions, or would like to work with me!

🤝 LinkedIn
🤝 juliespigner@gmail.com