Certainly! Based on the features available in the NMDB Aggregate Statistics Data Dictionary, here are some hypotheses that can be tested using various statistical tests:

1. T-Test (Comparing Means) Hypotheses:

- H1: The average loan amount for mortgages originated in the first half of a year is significantly different from those originated in the second half.

- **H2: The average property value of homes in coastal states is significantly different/lower from those in landlocked states.**

2. Chi-Square Test (Testing Independence) Hypotheses:

- H3: The distribution of loan origination numbers is independent of the year of origination.

**- H4: The distribution of average contract rates falls into distinct categories (e.g., low, medium, high) and is independent of the geographical region.**

3. ANOVA (Analysis of Variance) Hypotheses:

- H5: There's a significant difference in the average loan amounts across different regions (e.g., North, South, East, West).

- **H6: The average property value varies significantly across different interest rate categories (e.g., low, medium, high). → we have to bin the interest rate into categories.**

4. Correlation Test Hypotheses:

**- H7: There's a significant positive correlation between the average loan amount and the average property value.**

**- H8: There's a significant negative correlation between the average contract rate and the number of originations.**

5. Regression Analysis Hypotheses: → can be used in ML/DL models

- **H9: The average loan amount can be predicted using the average property value and average contract rate.**

**- H10: The number of originations can be predicted based on economic indicators (if external data is integrated).**

6. Non-parametric Tests (e.g., Mann-Whitney U Test) Hypotheses:

**- H11: The distribution of loan amounts for properties in urban areas is significantly different from those in rural areas.**

- **H12: The distribution of contract rates for mortgages originated in recession years is different from those in non-recession years.**

7. Time Series Analysis Hypotheses: (use for data analysis)

- H13: There's a seasonal pattern in the number of mortgage originations.

- H14: There's a trend in the average property value over the years.

8. Z-Test Hypotheses:

- H15: The average loan amount for a particular year is significantly different from the overall mean loan amount across all years.

- H16: The proportion of high-value loans (e.g., above $500,000) in a specific year is significantly different from the overall proportion across all years.

These hypotheses can serve as a starting point for your analysis. Depending on the results, you can further refine or expand your hypotheses. Always ensure that the chosen statistical test is appropriate for the type of data and the nature of the hypothesis.

Other possible hypotheses:

1. How mortgage rates, LTV, etc vary according to external factors such as employment rate, GDP.
2. Are these statistics like mortgage rates, LTV, related to demographic data (gender, race, ethnicity, age)?

Finding mortgage terms using classification/regression.

Tasks for ML/DL Model -

Apply both supervised and unsupervised learning techniques to perform classification, regression and pattern recognition. For example -

we can use regression use to predict average loan amount, average purchase price, average term to maturity, DTI ratio etc.

From the business’ perspective, we can use classification to predict the mortgage terms a client is likely to prefer adjustable rate mortgage or fixed rate mortgage. From the client perspective, we can predict whether being in his credit score range can get them high or low loan to value ratio etc.

Chosen hypotheses -

1. **The average property value of homes in coastal states is significantly different/lower from those in landlocked states.**
2. **The distribution of average contract rates falls into distinct categories (e.g., low, medium, high) and is independent of the geographical region.**
3. **The average property value varies significantly across different interest rate categories (e.g., low, medium, high.**
4. **There's a significant positive correlation between the average loan amount and the average property value.**
5. **There's a significant negative correlation between the average contract rate and the number of originations.**
6. **The average loan amount can be predicted using the average property value and average contract rate.**
7. **The number of originations can be predicted based on economic indicators (if external data is integrated).**
8. **The distribution of loan amounts for properties in urban areas is significantly different from those in rural areas.**
9. **The distribution of contract rates for mortgages originated in recession years is different from those in non-recession years.**
10. **How mortgage rates, LTV, etc vary according to external factors such as employment rate, GDP.**
11. **Are these statistics like mortgage rates, LTV, related to demographic data (gender, race, ethnicity, age)?**

**CHATGPT -**

Report Format

Introduction: In this analysis, we have formulated a series of hypotheses to explore various aspects of the mortgage and housing market. The goal is to gain insights into how different factors, both internal and external, impact key metrics and trends in the mortgage industry.

Hypotheses Testing:

Hypothesis 1: To investigate whether coastal states exhibit significant differences or lower average property values compared to landlocked states, we will conduct a comparative analysis of property values between these two categories of states. We will employ statistical tests to determine the significance of these differences.

Hypothesis 2: In assessing the distribution of average contract rates and their categorization into distinct groups, we will employ clustering and segmentation techniques. We will then analyze whether these categories are independent of geographical regions using statistical methods.

Hypothesis 3: To examine how the average property value varies across different interest rate categories, we will perform a comparative analysis of property values within these categories and assess their statistical significance.

Hypothesis 4: In order to establish whether there is a significant positive correlation between the average loan amount and the average property value, we will conduct a correlation analysis. This will help determine the strength and direction of the relationship between these variables.

Hypothesis 5: To investigate the potential negative correlation between the average contract rate and the number of originations, we will conduct a correlation analysis. This analysis will help us understand whether higher contract rates are associated with fewer mortgage originations.

Hypothesis 6: We will develop predictive models to test whether the average loan amount can be accurately predicted using the average property value and average contract rate as predictors. This will involve regression analysis and model evaluation.

Hypothesis 7: Exploring the predictability of the number of originations based on economic indicators will require integration of external data. We will construct predictive models using relevant economic indicators and assess their accuracy.

Hypothesis 8: To compare the distribution of loan amounts between urban and rural areas, we will conduct a comparative analysis and determine if there are significant differences in these distributions.

Hypothesis 9: We will assess whether the distribution of contract rates for mortgages varies between recession and non-recession years, utilizing appropriate statistical tests.

Hypothesis 10: Examining the relationship between mortgage rates, LTV ratios, and external factors such as employment rates and GDP will involve regression analysis and correlation tests.

Hypothesis 11: To explore the associations between mortgage-related statistics and demographic data (e.g., gender, race, ethnicity, age), we will conduct statistical analyses to identify any significant relationships or patterns.

Conclusion: Through rigorous statistical analyses and modeling, we aim to provide valuable insights into the mortgage and housing market, shedding light on the relationships between various factors and key mortgage-related metrics. This analysis will help stakeholders make informed decisions and better understand the dynamics of the mortgage industry.

Introduction:

We plan to explore various predictive tasks using Machine Learning (ML) models applied to the National Mortgage Database (NMDB®) Aggregate Mortgage Statistics. These tasks encompass classification, regression, and pattern recognition to gain valuable insights from the mortgage data. The objective is to enhance decision-making processes for both the mortgage business and clients.

Predictive Tasks:

1. Regression - Predicting Mortgage Metrics:

Average Loan Amount: We employ regression models to predict the average loan amount, helping lenders estimate the loan amounts borrowers are likely to request.

Average Purchase Price: Similar to loan amount, we predict the average purchase price, aiding in understanding property values.

Average Term to Maturity: Predicting the average term to maturity assists in assessing loan duration and risk.

Debt-to-Income (DTI) Ratio: Regression models can estimate borrowers' DTI ratios, aiding lenders in evaluating financial stability.

2. Classification - Mortgage Term Preference:

Mortgage Term Classification: Using classification models, we predict whether clients are more likely to prefer adjustable-rate mortgages (ARM) or fixed-rate mortgages (FRM). This can inform marketing and product offerings.

Loan-to-Value (LTV) Ratio Classification: We classify whether clients fall into high or low LTV ratio categories based on their credit scores, helping borrowers understand their potential down payment requirements.

3. Pattern Recognition - Identifying Trends:

Market Trend Recognition: Employing unsupervised learning techniques, we identify patterns and trends in mortgage data, helping businesses make informed decisions based on market dynamics.

Credit Score and LTV Trends: Uncovering trends in the relationship between credit scores and LTV ratios can aid in risk assessment and product development.

Additional Predictive Tasks:

4. Default Risk Prediction:

Using historical data, we can build models to predict the likelihood of loan defaults, assisting lenders in risk mitigation strategies.

5. Interest Rate Forecasting:

Predicting future interest rate trends using ML/DL models enables lenders to offer competitive rates and clients to make informed decisions.

6. Client Demographic Analysis:

We can use clustering to segment clients based on demographics, allowing businesses to tailor their services to specific client groups effectively.

7. Mortgage Approval Prediction:

Employing classification models, we can predict whether a mortgage application will be approved, streamlining the lending process.

8. Prepayment Prediction:

Predicting loan prepayments helps lenders manage cash flows and optimize mortgage portfolios.

Conclusion:

Leveraging ML models on NMDB® data offers a multitude of benefits, from enhancing decision-making processes to improving customer experiences. By predicting critical mortgage metrics, term preferences, and identifying market trends, businesses can adapt to changing market conditions. Moreover, these models can assist clients in making informed financial decisions. Additionally, tasks such as default risk prediction, interest rate forecasting, client demographics analysis, and mortgage approval prediction further enrich the mortgage industry's data-driven capabilities. Integrating ML/DL models into mortgage operations empowers stakeholders to make more precise, data-backed decisions, ultimately enhancing the efficiency and effectiveness of the mortgage market.

—-------------------------------

Problem Statement:

Understanding the complex interactions between mortgage metrics, demographic variables, geographic differences, and overarching external influences like GDP, unemployment rates, major world events like the COVID pandemic, and economic downturns becomes crucial in the dynamic housing market. Our goal is to begin a holistic examination of the mortgage and housing domain by utilizing the extensive data from the NMDB® Aggregate Mortgage Statistics.

With a focus on the intricacies of loan amounts, property valuations, and contractual rates, we intend to analyze the characteristics of the housing market. At the same time, we want to understand the complexities of demographics by examining the crucial roles that age, gender, ethnicity, and race play in mortgage dynamics. Another element of complication is provided by geography, which influences mortgage trends and property values due to regional variations.

But the housing market doesn't function alone. Mortgage behaviors are influenced by external macroeconomic factors like as GDP and unemployment rates, as well as rare occurrences like the COVID pandemic and economic crises. We plan to look at these externalities to determine how they affect and relate to our key mortgage metrics.

Our approach combines univariate and multivariate studies, enabling us to examine distinct aspects and how they relate to one another. We will be able to validate or disprove our initial hypotheses using rigorous statistical tests, which will clarify our investigations. Furthermore, the implementation of machine learning algorithms, both supervised and unsupervised, will improve our prediction powers while also giving housing market stakeholders relevant insights.

Our goal is to perform a comprehensive analysis of insights that can inform strategic planning, policy development, and decision-making in the housing market by fusing together the threads of mortgage metrics, demography, geography, and outside influences. We also focus on giving useful insights beneficial to both the business and its prospective clients.

Hypotheses:

1. Hypothesis 1 (H1): Coastal states have a significantly different average property value compared to landlocked states.

2. Hypothesis 2 (H2): The average contract rates can be categorized into distinct groups such as low, medium, and high, and these categories are not influenced by the geographical region.

3. Hypothesis 3 (H3): There is a significant variation in the average property value across different interest rate categories, namely low, medium, and high.

4. Hypothesis 4 (H4): There exists a strong positive correlation between the average loan amount and the average property value.

5. Hypothesis 5 (H5): A negative correlation is observed between the average contract rate and the number of mortgage originations.

6. Hypothesis 6 (H6): The average loan amount can be effectively predicted using variables such as the average property value and the average contract rate.

7. Hypothesis 7 (H7): Economic indicators, when integrated, can serve as reliable predictors for the number of mortgage originations.

8. Hypothesis 8 (H8): The distribution of loan amounts for properties significantly differs between urban and rural areas.

9. Hypothesis 9 (H9): Mortgages that originated during recession years exhibit a distinct distribution of contract rates compared to those that originated in non-recession years.

10. Hypothesis 10 (H10): Mortgage rates, LTV, and other related metrics vary in accordance with external factors such as the employment rate and GDP.

11. Hypothesis 11 (H11): Key mortgage statistics, including mortgage rates and LTV, exhibit patterns or relationships with demographic data points such as gender, race, ethnicity, and age.

Note: As our project progresses, we anticipate the possibility of uncovering new insights and patterns in the data. This may lead us to formulate additional hypotheses for testing, ensuring a comprehensive and in-depth analysis of the mortgage and housing landscape.

---

Machine Learning Exploration on NMDB® Aggregate Mortgage Statistics

We aim to deploy Machine Learning (ML) models to delve deep into the National Mortgage Database (NMDB®) Aggregate Mortgage Statistics. Through a blend of classification, regression, and pattern recognition techniques, we aim to unearth valuable insights that can guide both mortgage businesses and potential clients.

Predictive Tasks:

1. Regression Tasks (supervised): Models → Linear Regression, Random Forest, ANN

- **Average Loan Amount Prediction**: Our goal is to use regression models to forecast the average loan amount. This will provide lenders with a clearer picture of the potential loan amounts borrowers might seek.

- **Average Purchase Price Estimation**: By predicting the average purchase price, we aim to offer insights into prevailing property values.

- **Debt-to-Income (DTI) Ratio Forecasting**: Estimating the DTI ratio of borrowers will assist lenders in gauging the financial health of their clients.

- **Interest Rate Trend Prediction**: With ML models, we'll forecast potential interest rate movements, enabling lenders to position their rates competitively and aiding clients in their decision-making process.

2. Classification Tasks:

- Mortgage Term Preference: We'll deploy classification models to determine if borrowers have a leaning towards adjustable-rate mortgages (ARM) or fixed-rate mortgages (FRM). This insight can shape marketing strategies and product design.

- Loan-to-Value (LTV) Ratio Classification: By classifying clients into high or low LTV ratio brackets based on their credit scores, we aim to guide borrowers about potential down payment expectations.

3. Pattern Recognition: Model → Clustering

- Market Trend Analysis: Using unsupervised learning, we'll spot patterns and trends in the mortgage landscape, offering businesses a roadmap to navigate market shifts.

- Credit Score and LTV Relationship: We'll explore the trends between credit scores and LTV ratios, providing valuable data for risk assessment and product tailoring.

- **Client Demographic Segmentation**: Through clustering techniques, we'll segment clients based on demographics, allowing businesses to customize their offerings for diverse client groups.

Conclusion:

By applying ML models to the NMDB® dataset, we stand to gain a plethora of insights that can redefine decision-making processes and elevate client experiences. From predicting pivotal mortgage metrics to discerning market trends, our analyses will equip businesses to stay agile in a fluctuating market. Furthermore, our models will serve as a beacon for clients, guiding them toward informed financial choices. As we integrate these models into the mortgage ecosystem, we're not just enhancing operational efficiency but also championing a data-driven approach in the mortgage industry.

Note: The tasks mentioned above are based on an initial analysis of the features available in the dataset. As the project progresses, the feasibility of each task will be evaluated in-depth, and adjustments will be made accordingly.

New work:

While previous reports, such as the one on first-time homebuyers, touched upon demographic and geographic dimensions of the mortgage landscape, our study takes a distinct approach by intertwining these metrics with external economic factors. We will try to integrate key indicators like GDP, unemployment rates, and economic crises to provide a richer context to mortgage behaviors, an angle not deeply explored in the prior reports.

Our analysis stands out in its rigorous multivariate approach, examining the interplay between multiple variables simultaneously, offering a more comprehensive view of the mortgage ecosystem. Through hypothesis testing using statistical measures, we've not only posed questions but also plan to rigorously validate them, ensuring our insights are both novel and robust.

Furthermore, we also aim to deploy various ML modeling techniques to derive better insights from the data, something that has not been done in previous works. This combination of economic context, rigorous statistical validation, and predictive modeling ensures our analysis offers fresh perspectives, filling gaps left by previous studies.

SOURCE : 0

FREQUENCY : 0

GEOLEVEL : 0

GEOID : 0

GEONAME : 0

MARKET : 0

PERIOD : 0

YEAR : 0

QUARTER : 0

MONTH : 0

SUPPRESSED : 0

TOT\_ORIG : 73

AVE\_LOANAMT : 73

AVE\_PROPVAL : 73

AVE\_INTRATE : 73

PCT\_INTRATE\_LT\_3 : 11

PCT\_INTRATE\_3\_4 : 11

PCT\_INTRATE\_4\_5 : 11

PCT\_INTRATE\_5\_6 : 11

PCT\_INTRATE\_GE\_6 : 11

PCT\_OWNOCC : 73

PCT\_FTHB : 9421

PCT\_REPEATHB : 9421

PCT\_HP : 73

PCT\_CASHOUT : 73

PCT\_OTH\_REFI : 73

PCT\_REFI : 73

AVE\_TERM : 73

PCT\_ARM : 73

PCT\_TERM\_FRM\_15 : 73

PCT\_TERM\_FRM\_30 : 73

AVE\_DTI : 73

PCT\_DTI\_LE36 : 73

PCT\_DTI\_3743 : 73

PCT\_DTI\_GE44 : 73

AVE\_VANTAGESCR : 73

PCT\_VS\_VERYPOOR : 73

PCT\_VS\_POOR : 73

PCT\_VS\_FAIR : 73

PCT\_VS\_GOOD : 73

PCT\_VS\_EXCELLENT : 73

AVE\_LTV : 73

AVE\_CLTV : 73

PCT\_LTV\_LE70 : 73

PCT\_LTV\_7080 : 73

PCT\_LTV\_8090 : 73

PCT\_LTV\_9095 : 73

PCT\_LTV\_9597 : 73

PCT\_LTV\_GT97 : 73

PCT\_GOVERNMENT : 73

PCT\_ENTERPRISE : 73

PCT\_OTHERCONFORMING : 73

PCT\_NONCONFORMING : 73

PCT\_WHT : 9421

PCT\_BLK : 9421

PCT\_ASN : 9421

PCT\_HPI : 9421

PCT\_AMI : 9421

PCT\_MIX : 9421

PCT\_HIS : 9421

PCT\_HSP : 9421

PCT\_WNH : 9421

PCT\_MNH : 9421

AVE\_AGE\_BORROWER : 73

PCT\_AGE\_LT25 : 73

PCT\_AGE\_2534 : 73

PCT\_AGE\_3544 : 73

PCT\_AGE\_4554 : 73

PCT\_AGE\_5564 : 73

PCT\_AGE\_GE65 : 73

PCT\_MALEBOR : 73

PCT\_FEMALEBOR : 73

PCT\_TWOBOR : 73

PCT\_MULTIBOR : 73

Filling with average value:

data['AVE\_LOANAMT'] = data['AVE\_LOANAMT'].fillna(data['AVE\_LOANAMT'].mean())

data['AVE\_PROPVAL'] = data['AVE\_PROPVAL'].fillna(data['AVE\_PROPVAL'].mean())

data['AVE\_INTRATE'] = data['AVE\_INTRATE'].fillna(data['AVE\_INTRATE'].mean())

data['TOT\_ORIG'] = data['TOT\_ORIG'].fillna(data['TOT\_ORIG'].mean())

Filling with proportion:

#fill missing values in PCT\_VS\_VERYPOOR, PCT\_VS\_POOR, PCT\_VS\_FAIR, PCT\_VS\_GOOD, PCT\_VS\_EXCELLENT with their ratio to the sum of these columns

data['PCT\_VS\_VERYPOOR'] = data['PCT\_VS\_VERYPOOR'].fillna((data['PCT\_VS\_VERYPOOR'].mean() \* 100)/(data['PCT\_VS\_VERYPOOR'].mean()+data['PCT\_VS\_POOR'].mean()+data['PCT\_VS\_FAIR'].mean()+data['PCT\_VS\_GOOD'].mean()+data['PCT\_VS\_EXCELLENT'].mean()))

data['PCT\_VS\_POOR'] = data['PCT\_VS\_POOR'].fillna((data['PCT\_VS\_POOR'].mean() \* 100)/(data['PCT\_VS\_VERYPOOR'].mean()+data['PCT\_VS\_POOR'].mean()+data['PCT\_VS\_FAIR'].mean()+data['PCT\_VS\_GOOD'].mean()+data['PCT\_VS\_EXCELLENT'].mean()))

data['PCT\_VS\_FAIR'] = data['PCT\_VS\_FAIR'].fillna((data['PCT\_VS\_FAIR'].mean() \* 100)/(data['PCT\_VS\_VERYPOOR'].mean()+data['PCT\_VS\_POOR'].mean()+data['PCT\_VS\_FAIR'].mean()+data['PCT\_VS\_GOOD'].mean()+data['PCT\_VS\_EXCELLENT'].mean()))

data['PCT\_VS\_GOOD'] = data['PCT\_VS\_GOOD'].fillna((data['PCT\_VS\_GOOD'].mean() \* 100)/(data['PCT\_VS\_VERYPOOR'].mean()+data['PCT\_VS\_POOR'].mean()+data['PCT\_VS\_FAIR'].mean()+data['PCT\_VS\_GOOD'].mean()+data['PCT\_VS\_EXCELLENT'].mean()))

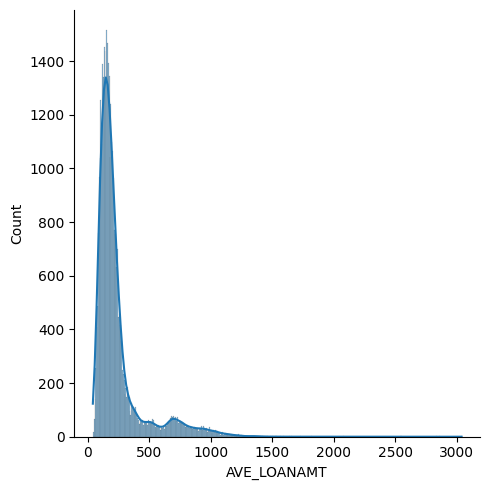
data['PCT\_VS\_EXCELLENT'] = data['PCT\_VS\_EXCELLENT'].fillna((data['PCT\_VS\_EXCELLENT'].mean() \* 100)/(data['PCT\_VS\_VERYPOOR'].mean()+data['PCT\_VS\_POOR'].mean()+data['PCT\_VS\_FAIR'].mean()+data['PCT\_VS\_GOOD'].mean()+data['PCT\_VS\_EXCELLENT'].mean()))

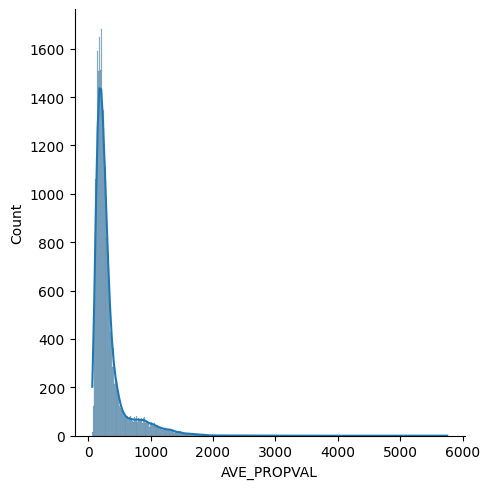
Columns deleted as they were not relevant:

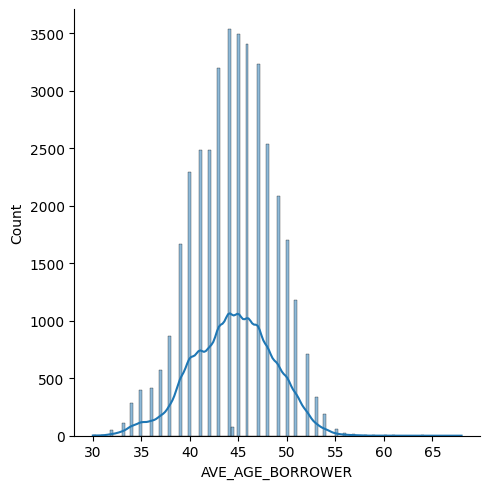
'SOURCE', 'FREQUENCY', 'QUARTER', 'MONTH', 'SUPPRESSED', 'PERIOD'

Columns deleted as they were not possible to impute:

'PCT\_FTHB', 'PCT\_REPEATHB', 'PCT\_HP', 'PCT\_CASHOUT', 'PCT\_OTH\_REFI', 'PCT\_REFI'







AVE\_LOANAMT skewness: 2.776116846197745

AVE\_LOANAMT kurtosis: 10.295564840229021

AVE\_PROPVAL skewness: 3.0851284285985847

AVE\_PROPVAL kurtosis: 15.663047136875

AVE\_INTRATE skewness: 0.26261490595231524

AVE\_INTRATE kurtosis: -1.0809840060528428

AVE\_TERM skewness: -0.9476874380552185

AVE\_TERM kurtosis: 1.036102395321957

AVE\_DTI skewness: 0.362830346537872

AVE\_DTI kurtosis: 0.39184970068256275

AVE\_VANTAGESCR skewness: -0.46053957460588196

AVE\_VANTAGESCR kurtosis: -0.44048745393896294

AVE\_LTV skewness: 0.44611440618150994

AVE\_LTV kurtosis: 0.07354075256105519

AVE\_CLTV skewness: 0.3433746612833633

AVE\_CLTV kurtosis: -0.008295052452037321

AVE\_AGE\_BORROWER skewness: -0.15948758263132395

AVE\_AGE\_BORROWER kurtosis: -0.09804880413307693