Loan Approval Analysis

DSS Final Project

Batsheva Levin

Introduction

In this project, we analyze a loan approval dataset to understand the key factors influencing loan decisions. Using exploratory data analysis, statistical methods, and machine learning, we identify patterns, uncover insights, and build predictive models to improve decision-making. This presentation walks through the approach, findings, and recommendations for optimizing the loan approval process.

Exploratory Data Analysis

EDA Step 1: Data Collection and Understanding

- We are working with the loan approval dataset from Kaggle
- Load the dataset into Python (Pandas data frame) and Tableau for analysis
- The dataset contains 4,269 rows and 13 columns
- Variables include loan amount, loan term in years, credit score, annual income, number of dependents, etc.

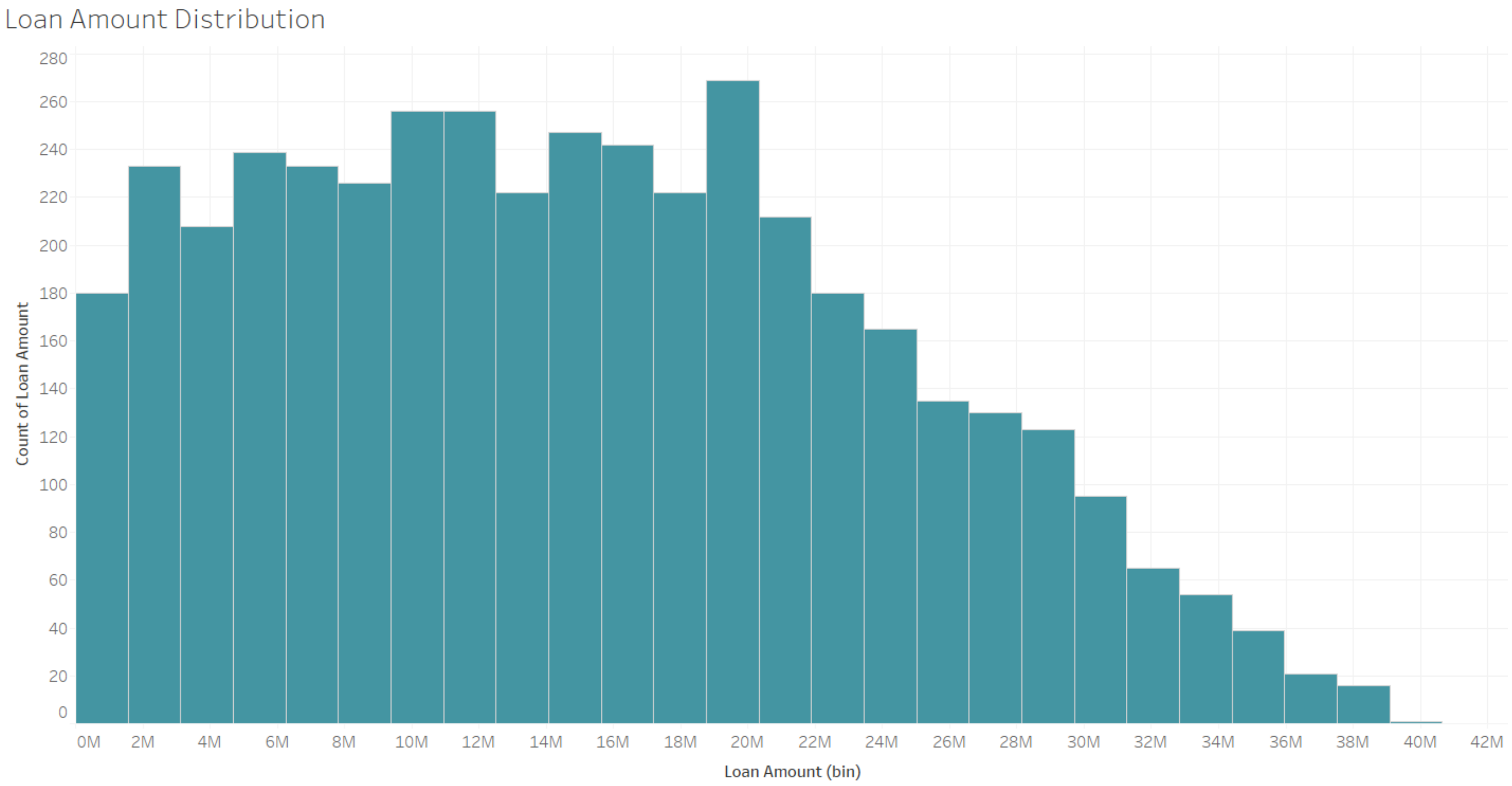
EDA Step 2: Data Cleaning and Preparation

- Worked in Python and assured that there are no null values or duplicates
- Ensured correct data types (numerical values are stored as integers) and made changes accordingly

EDA Step 3: Visualize Data Distributions

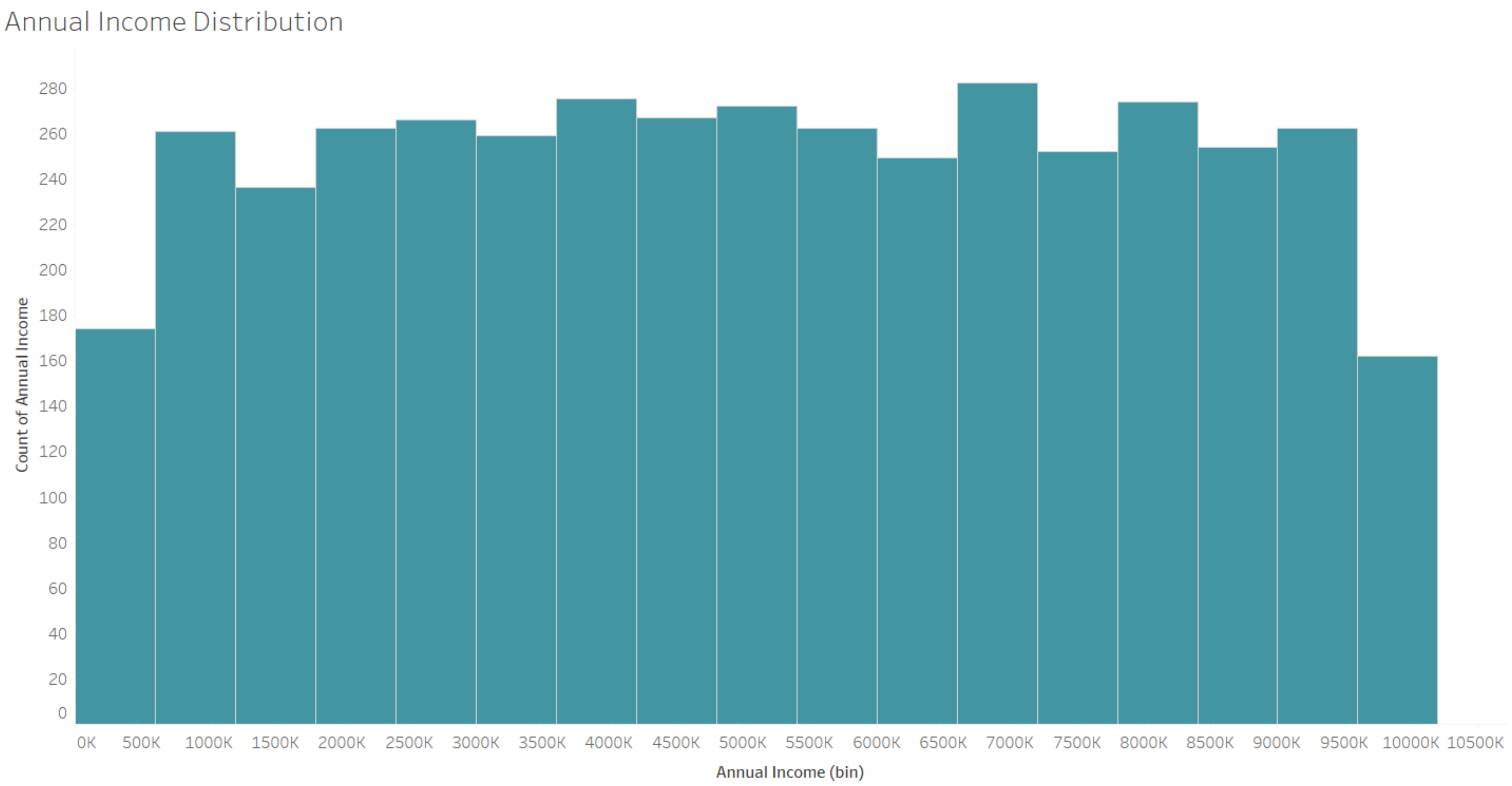
- Created multiple charts in Tableau to analyze the distributions of different variables

- Loan amount requests range from under $2 million - $40 million
- The distribution is slightly right skewed



Loan Amount Distribution

- Annual income of applicants ranges from under $500K -$1 million
- Distribution is nearly uniform



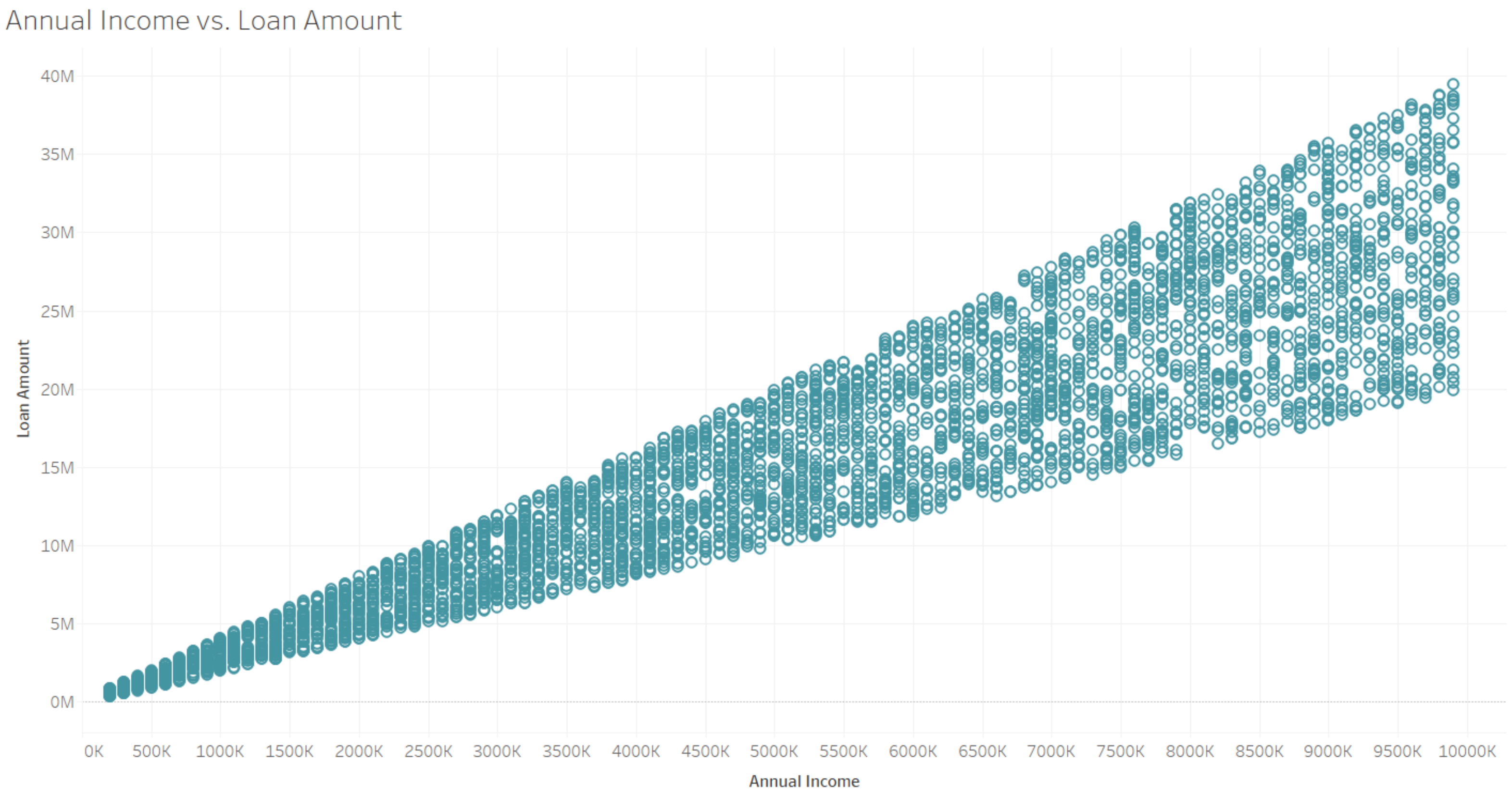
Annual Income Distribution

EDA Step 4: Identify Patterns and Trends

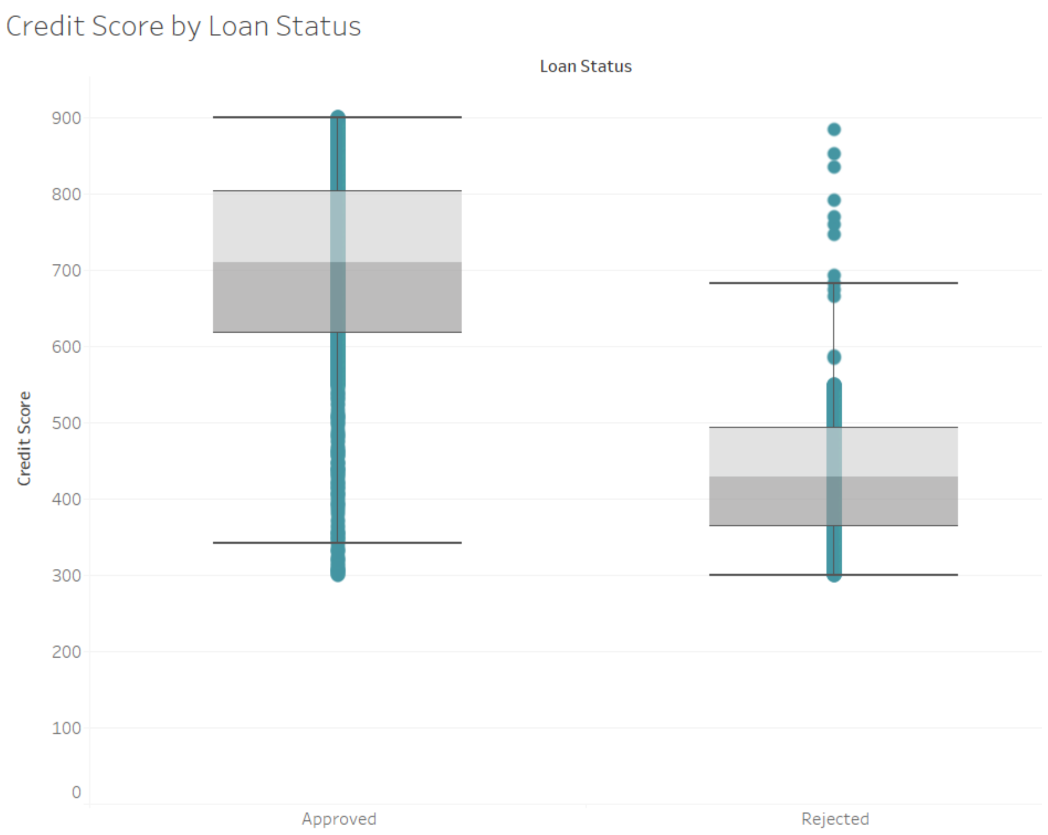
- Analyzed relationships between variables and discovered patterns
- Examined correlation to detect strong associations
- Further details regarding these charts will be provided in the steps of descriptive analytics

Annual Income vs. Loan Amount

- There is a strong positive correlation between income and requested loan amount
- These variables have a correlation of .93
- It is logical that individuals with higher incomes are likely to request larger loan amounts.

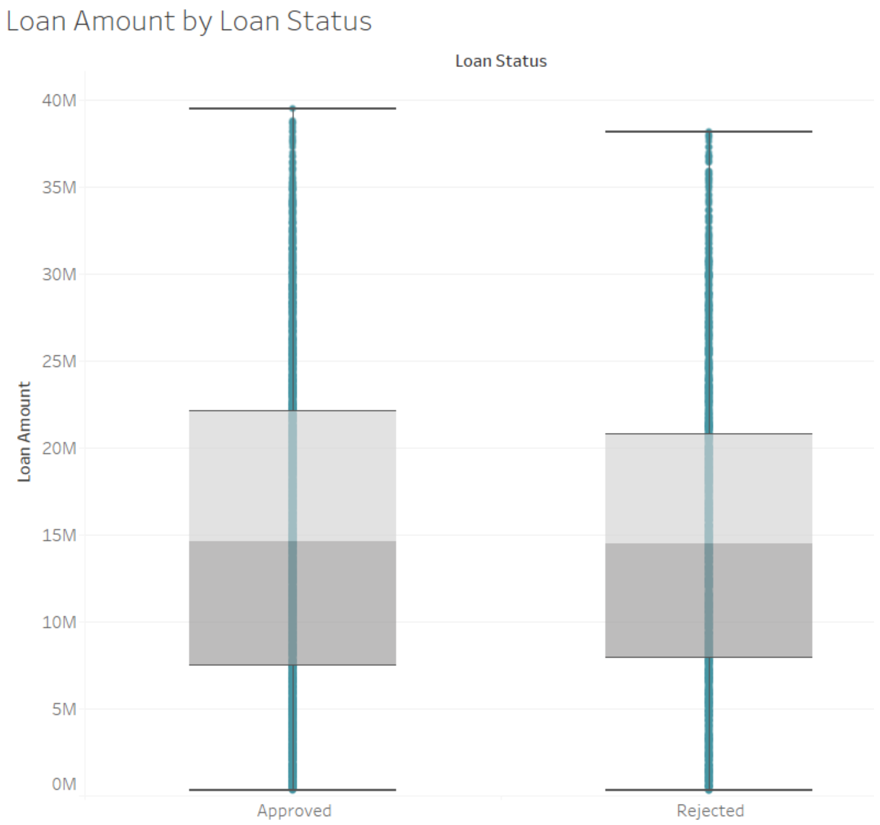


- Rejected loans have a much lower mean credit score.
- The graph for approved seems to be skewed, with many data points lying below the mean.
- The rejected loans have some outliers with high credit scores.



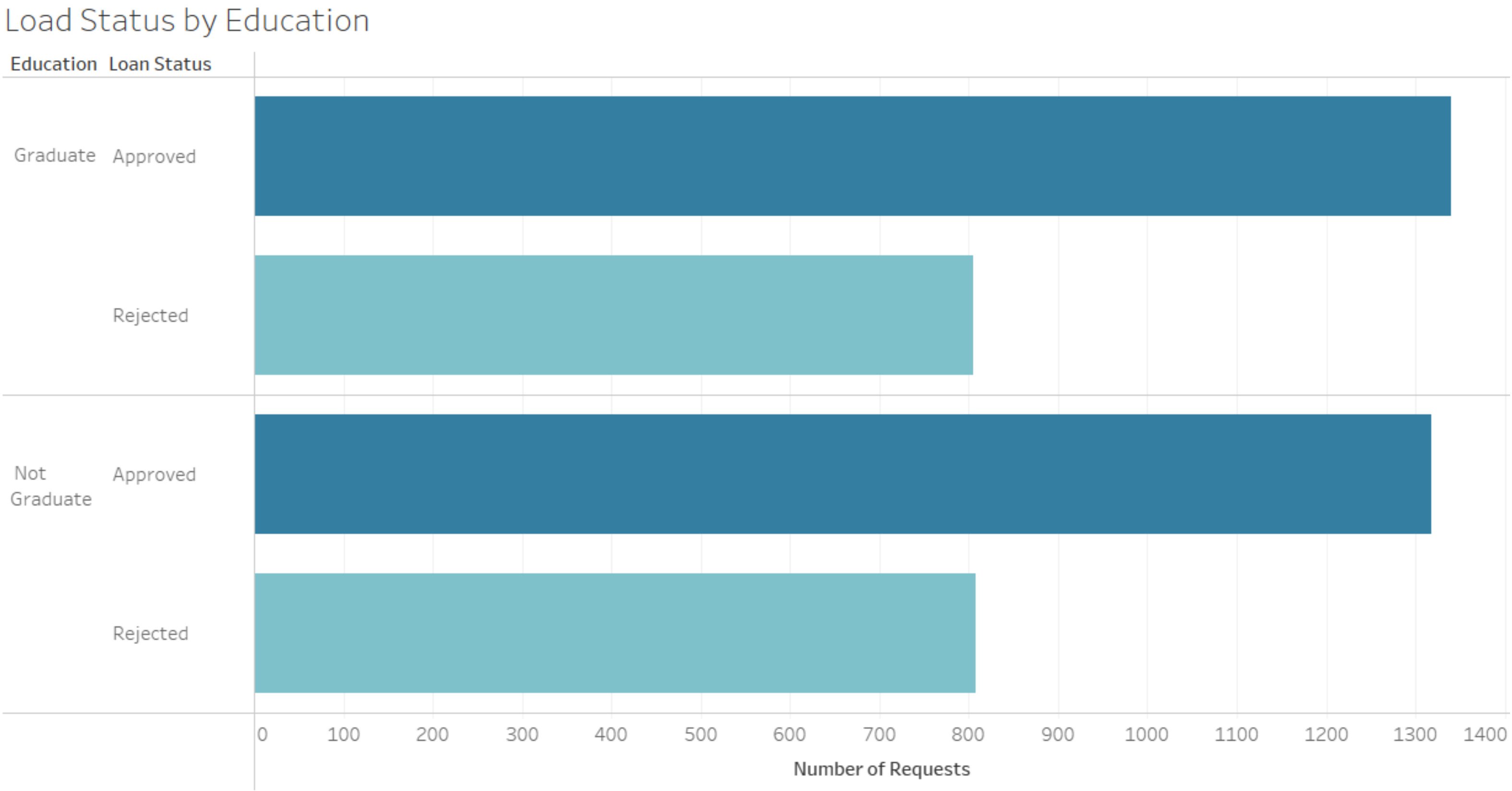
Credit Score by Loan Status

- Loan amounts seem to be similar for approved and rejected loans



Loan Amount by Loan Status

- The number of approved/rejected loans does not seem to change significantly based on an applicant's education

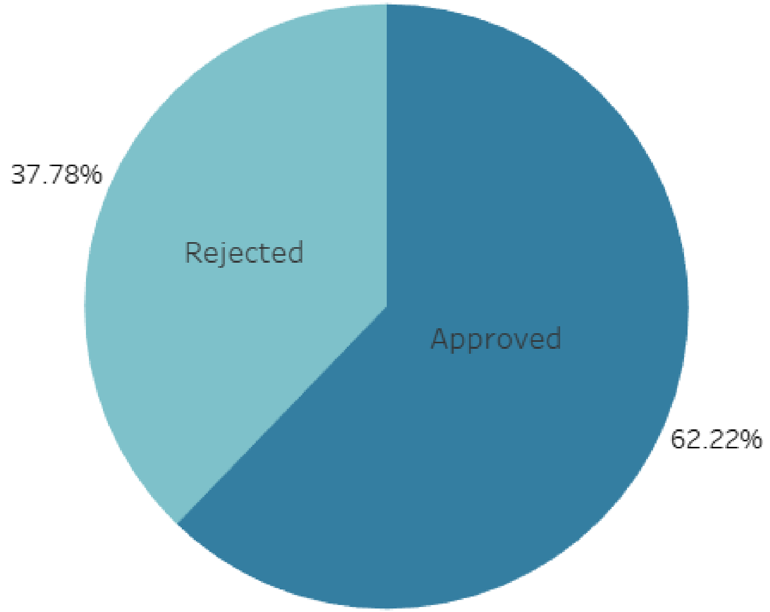


Loan Status by Education

Descriptive Analytics

Looks at past data to provide insights into what has happened

How many loans were approved/rejected?



- Approved: 62.22% (2,656 loans)
- Rejected: 37.78% (1,613 loans)

Summary Statistics

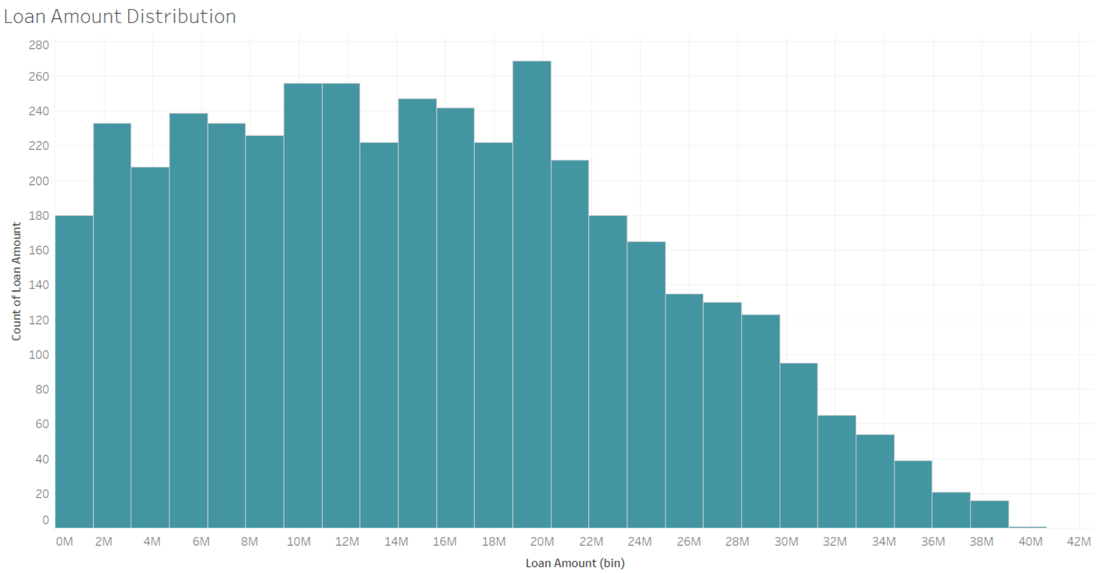
```
data.describe()
```

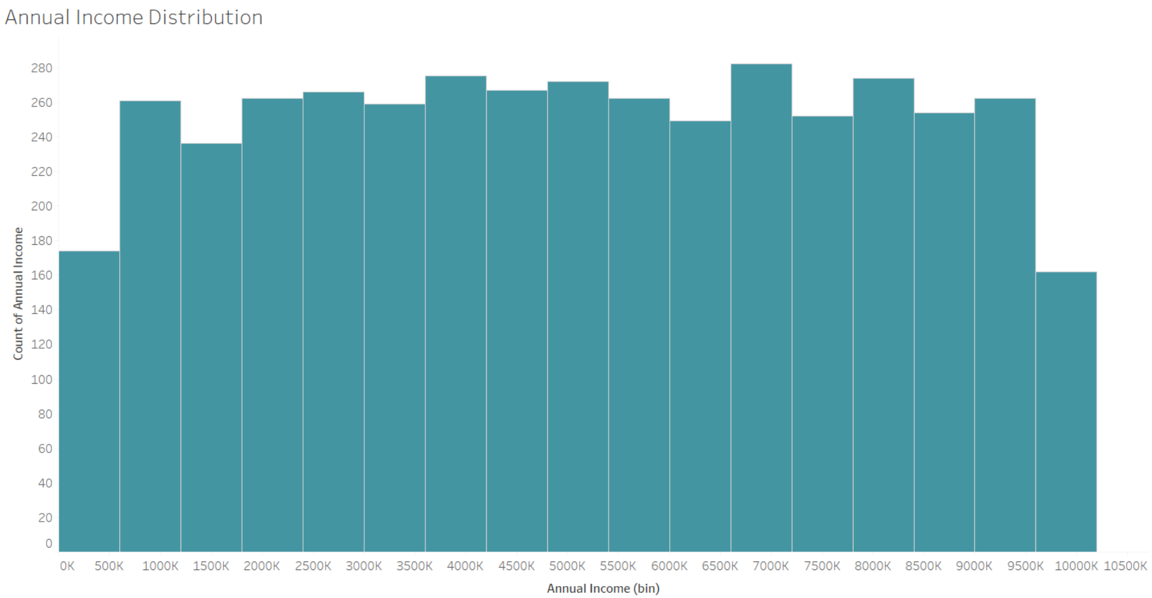
	loan_id	no_of_dependents	loan_term_years	credit_score	residential_assets_value	commercial_assets_value	luxury_assets_value	bank_asset_value
count	4269.000000	4269.000000	4269.000000	4269.000000	4.269000e+03	4.269000e+03	4.269000e+03	4.269000e+03
mean	2135.000000	2.498712	10.900445	599.936051	7.472617e+06	4.973155e+06	1.512631e+07	4.976692e+06
std	1232.498479	1.695910	5.709187	172.430401	6.503637e+06	4.388966e+06	9.103754e+06	3.250185e+06
min	1.000000	0.000000	2.000000	300.000000	-1.000000e+05	0.000000e+00	3.000000e+05	0.000000e+00
25%	1068.000000	1.000000	6.000000	453.000000	2.200000e+06	1.300000e+06	7.500000e+06	2.300000e+06
50%	2135.000000	3.000000	10.000000	600.000000	5.600000e+06	3.700000e+06	1.460000e+07	4.600000e+06
75%	3202.000000	4.000000	16.000000	748.000000	1.130000e+07	7.600000e+06	2.170000e+07	7.100000e+06
max	4269.000000	5.000000	20.000000	900.000000	2.910000e+07	1.940000e+07	3.920000e+07	1.470000e+07

Summary Statistics

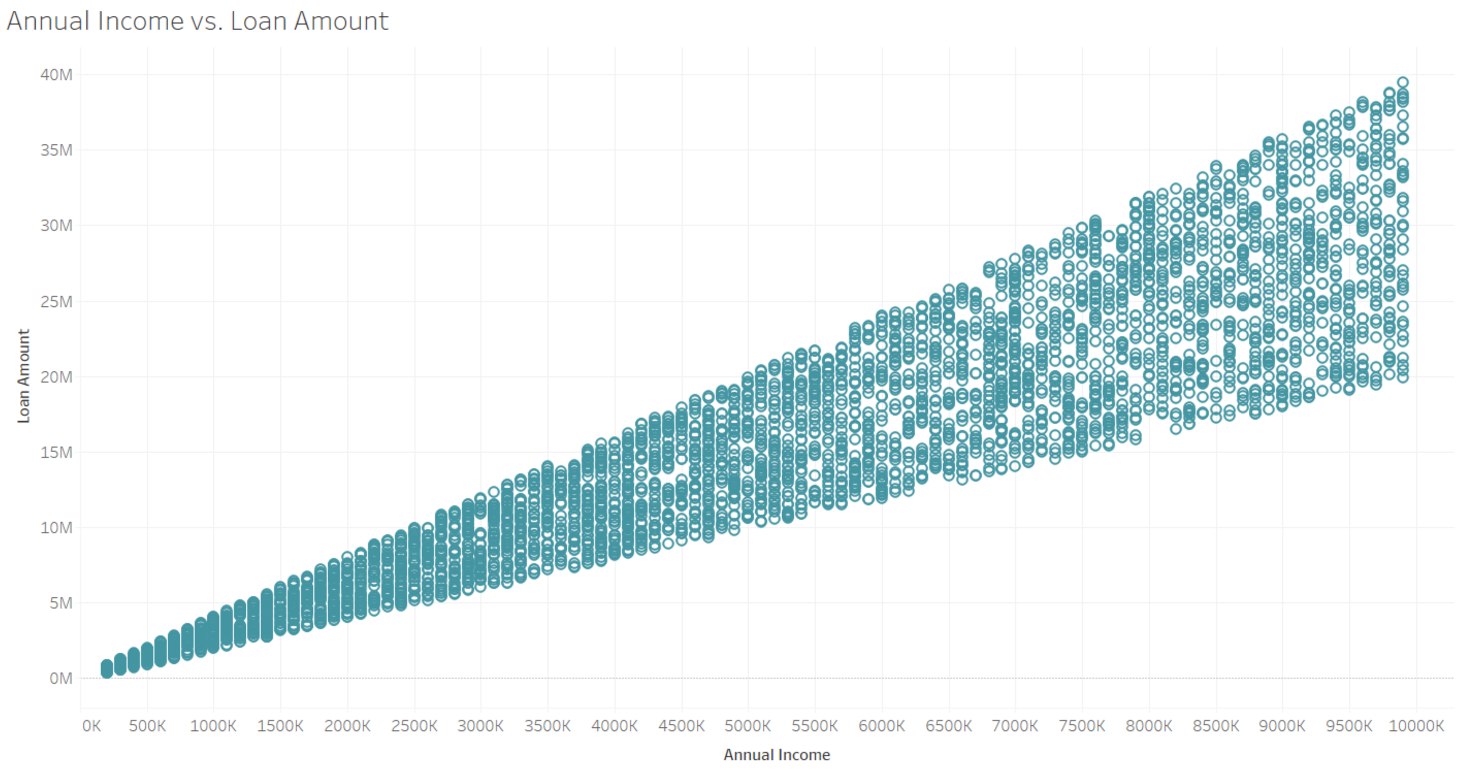
- Average loan amount by loan status:
  - Approved: $15,247,250
  - Rejected: $14,946,060
- Average credit score by loan status
  - Approved: 703
  - Rejected: 429
- Average number of dependents:
  - Approved: 2.47
  - Rejected: 2.54

Data Distributions





Correlation



Categorical Analysis

- Percentage of applicants that are self employed
  - Yes: 50.36%
  - No: 49.64
- Percentage of applicants that are graduates
  - Yes: 50.22%
  - No: 49:78%

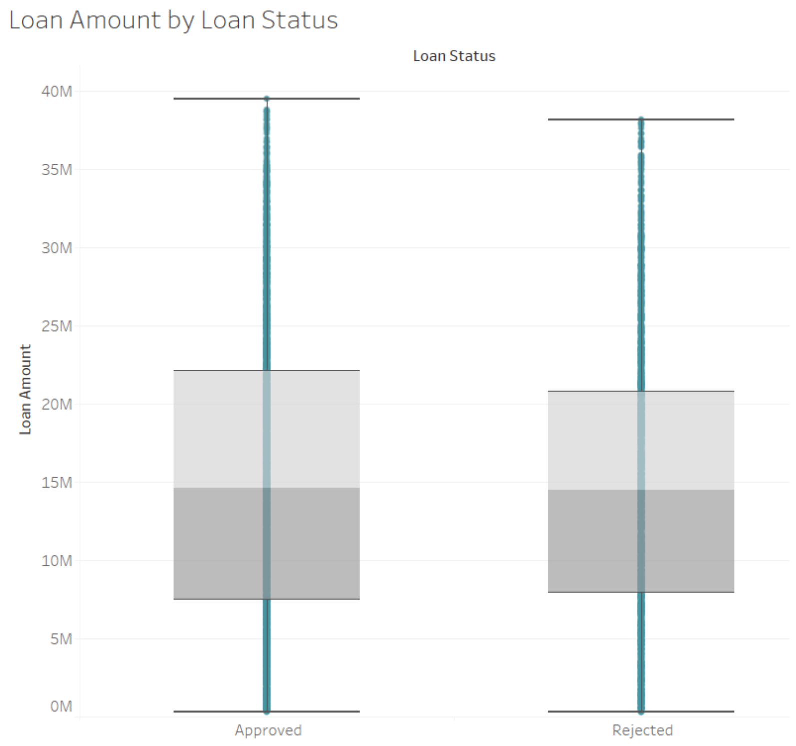
Diagnostic Analytics

Examines data to understand why something happened by identifying patterns and anomalies

Why do loans get approved?

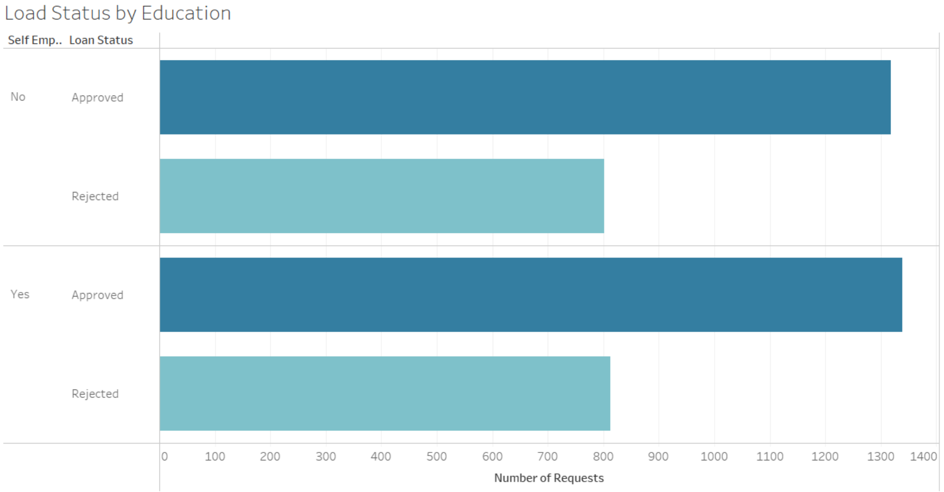
Are lower loan amounts more likely to get approved?

- Average loan amount by loan status:
  - Approved: $15,247,250
  - Rejected: $14,946,060
- Loan amounts are similar for approved and rejected loans
- Smaller loans do not guarantee a higher likelihood of approval



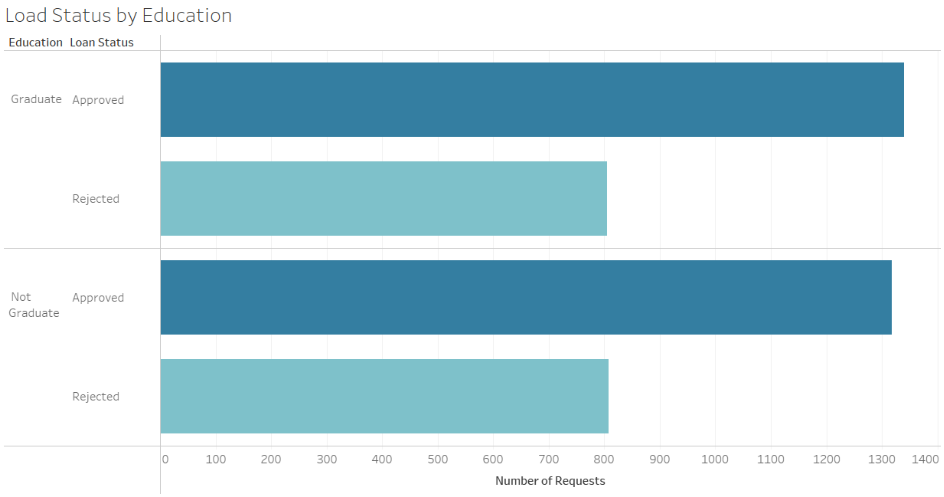
Do salaried applicants get approved more often?

- Of the approved loans, 1,338 were self-employed and 1,318 were not.
- Employment status does not appear to impact loan approval significantly



Are graduates more likely to be approved for loans?

- Of the approved loans, 1,339 were graduates and 1,317 were not.
- Education does not appear to impact loan approval significantly



Is loan approval connected to the number of dependents?

- The average number of dependents for both approved and rejected loans is nearly identical, indicating that the number of dependents does not significantly influence loan approval decisions.

```
avg_loan_amount = data.groupby(" loan_status")[" no_of_dependents"].mean()
avg_loan_amount
```

```
 loan_status
 Approved     2.474774
 Rejected     2.538128
Name:  no_of_dependents, dtype: float64
```

Are shorter-term loans more likely to be approved?

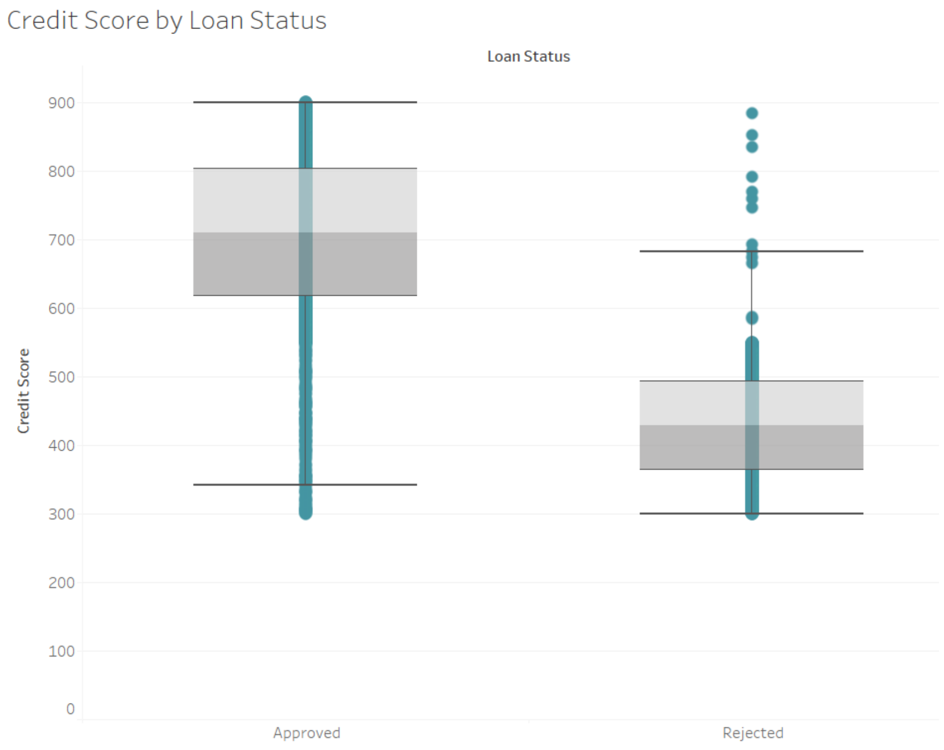
- The average loan term for approved loans was 10.4 years, while rejected loans had an average term of 11.7 years.
- There is only a slight difference in the number of years for approved and rejected loans.

```
avg_loan_amount = data.groupby(" loan_status")[" loan_term_years"].mean()
avg_loan_amount
```

```
 loan_status
 Approved     10.397590
 Rejected     11.728456
Name:  loan_term_years, dtype: float64
```

Are most approved loans linked to higher credit scores?

- The average credit scores for approved and rejected loans are drastically different. Approved loans have an average credit score of 703, and rejected loans have an average of 429.
- However, the boxplots are skewed, indicating that credit score may not be the only factor involved in assessing loan approvals.



Does credit score outweigh income when in comes to loan approvals?

- Low-credit applicants get rejected regardless of income, indicating that credit scores hold more significance than income.

Conclusion of Diagnostic Analytics

Credit score seems to be the strongest predictor of loan approval. Other factors such as income, education, and dependents show no significant impact. There may be additional hidden factors (such as employment history, previous loans, or bank policies) that are affecting loan approval. These variables can be areas for future investigation.

Predictive Analytics

Uses historical data and statistical models to forecast future outcomes

Logistic Regression Model

Logistic Regression Model

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report
```

```
data[' loan_status'] = data[' loan_status'].replace({'Approved': 1, 'Rejected': 0})
```

```
X = data[['credit_score', 'annual_income', ' loan_amount', ' loan_term_years']]
y = data[' loan_status']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
model = LogisticRegression()
model.fit(X_train, y_train)
```

```
▾ LogisticRegression
LogisticRegression()
```

```
y_pred = model.predict(X_test)
```

Evaluate Accuracy

This model has an accuracy of .91, meaning the model correctly predicts loan approval/rejection 91% of the time.

```
print(f"Accuracy: {accuracy_score(y_test, y_pred):.2f}")
```

```
Accuracy: 0.91
```

Confusion Matrix

- A confusion matrix is a table that assesses the performance of a classification model by comparing its predictions to the actual values, providing a detailed breakdown of correct and incorrect classifications.
  - 497 true negatives: the model correctly predicted rejected loans as rejected
  - 39 false positives: the model incorrectly predicted approved loans as rejected
  - 42 false negatives: the model incorrectly predicted rejected loans as approved
  - 276 true positives: the model correctly predicted approved loans as approved

```
Confusion Matrix:
[[497  39]
 [ 42 276]]
```

Classification Report

```
print(f"Classification Report:\n{classification_report(y_test, y_pred)}")
```

```
Classification Report:
              precision    recall  f1-score   support

    Approved       0.92      0.93      0.92       536
    Rejected       0.88      0.87      0.87       318

    accuracy                           0.91       854
   macro avg       0.90      0.90      0.90       854
weighted avg       0.90      0.91      0.91       854
```

Model Adjustment

- Increasing precision/minimizing false positive may be desired by the bank to avoid approving applicants who may not pay back their loans.
- The regression model was modified, but the adjustments greatly reduced accuracy, precision, and recall. As a result, the original model was retained.

Impact of Features on Loan Approval

Credit score has the strongest positive impact

```
feature_importance = pd.DataFrame({'Feature': X.columns, 'Coefficient': abs(model.coef_[0])})
feature_importance.sort_values(by='Coefficient', ascending=False)
```

	Feature	Coefficient
0	credit_score	4.154630
2	loan_amount	1.274905
1	annual_income	1.204029
3	loan_term_years	0.854564

Prescriptive Analytics

Suggests actions based on previous analysis to optimize outcomes and improve decision-making

Prescriptive Analytics: Recommendation #1

Credit scores are a critical factor in evaluating loan applications. It is highly recommended that the bank prioritize credit score assessments when determining loan approval.

Prescriptive Analytics: Recommendation #2

- Hidden factors may exist outside the current dataset. It is recommended that the bank conduct further investigations and collect additional data to uncover insights that could impact loan approval decisions.

- Possible hidden factors can include:
  - Financial behavior
  - Past loan history
  - Purpose of the loan
  - Employment factors