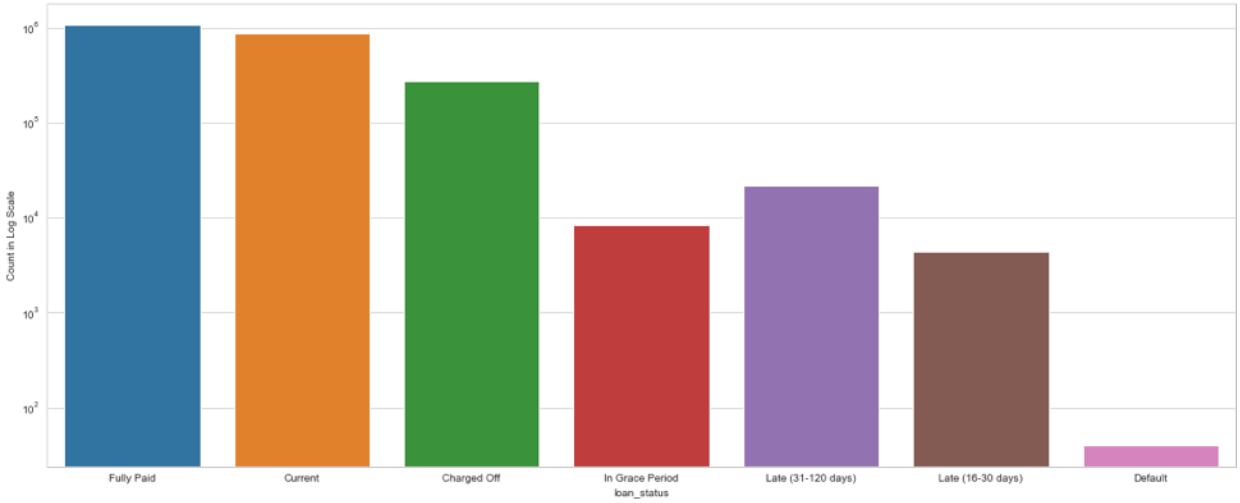
In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
from myEda import EDA #my eda python file

%matplotlib inline

acc_filepath = './archive/accepted_2007_to_2018q4.csv/'
rej_filepath = './archive/rejected_2007_to_2018q4.csv/'
data = pd.read_csv(acc_filepath+"accepted_2007_to_2018q4.csv")
eda = EDA(data)
eda.preprocessing()
```

Note:

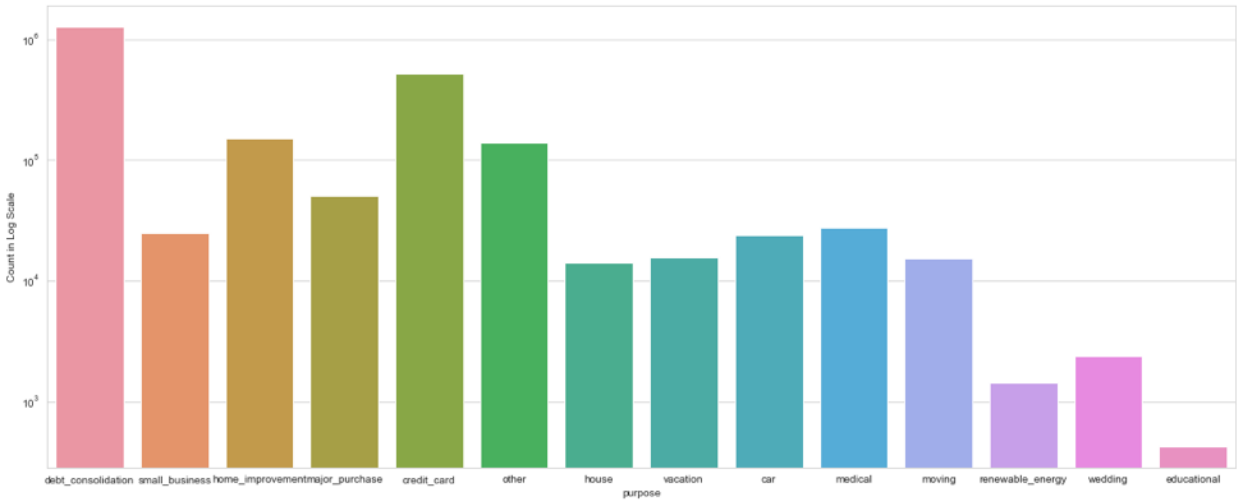
- If the maps are not showing in IPYNB then please check the pdf. The notebook needs to be executed for the maps plots to show.
- All observations are summarized at the end of the notebook.

Accepted csv eda

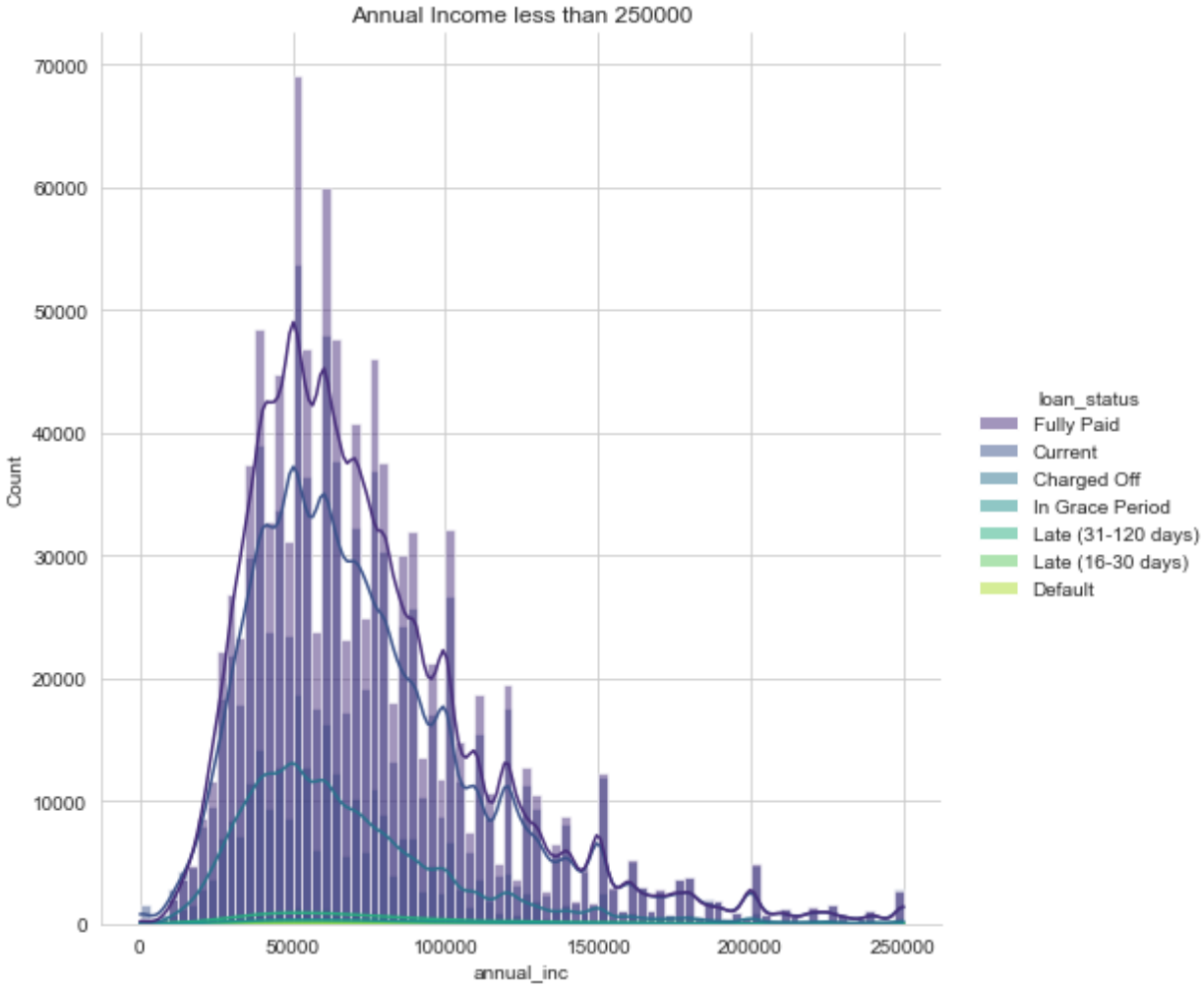
In [2]:
```
eda.logcountplot(x="loan_status")
```



In [3]:
```
1 eda.logcountplot("purpose")
```



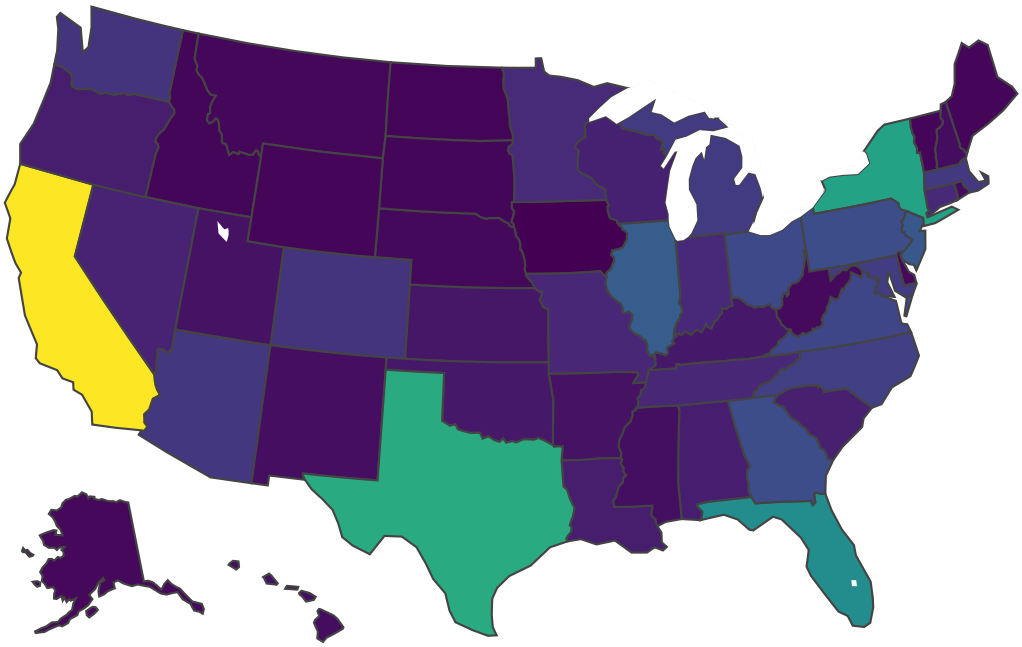
In [4]:
```
1 eda.incomedist()
```



Total Loan

In [5]:
```
eda.mapdist_loan_amt()
```

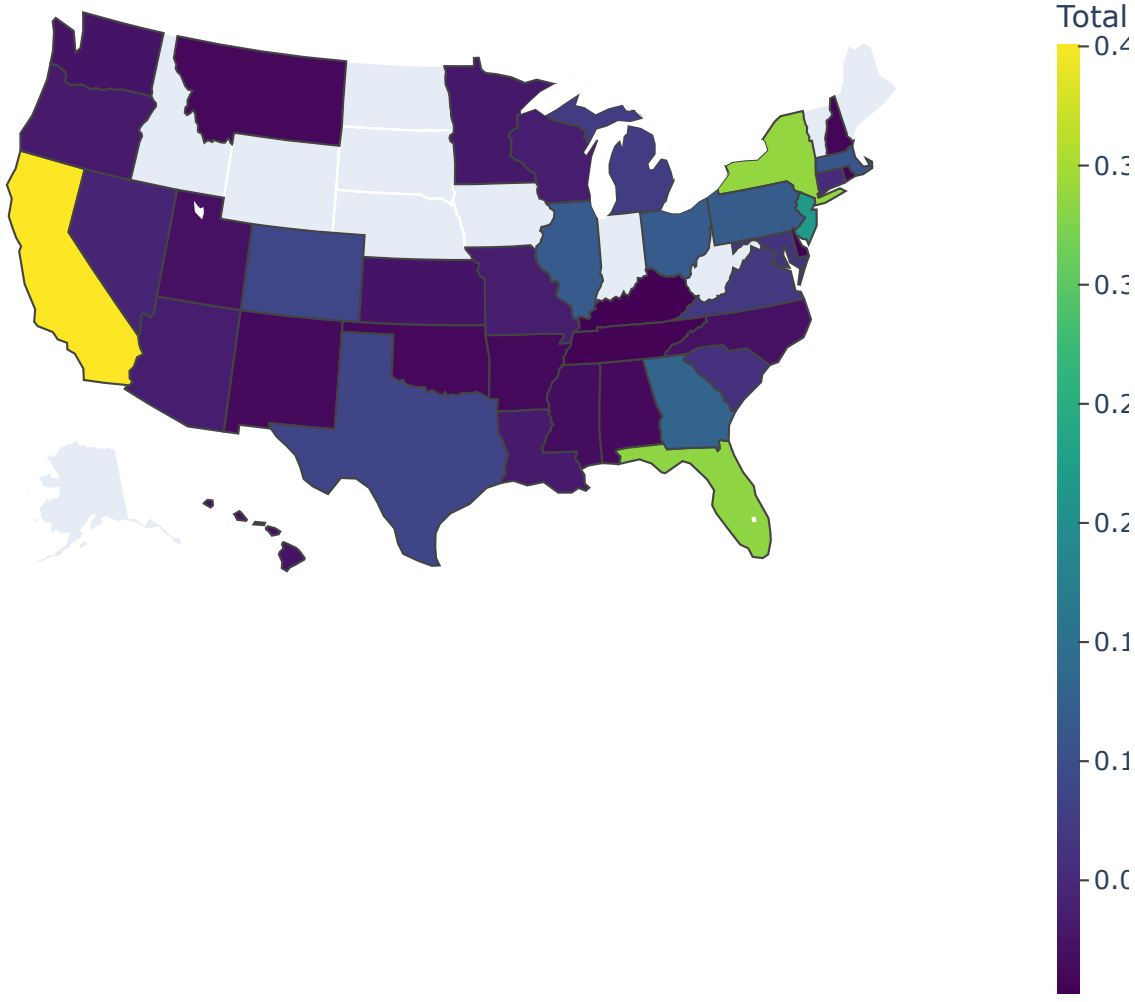
In [26]:
```
%%html
<iframe src="./choropleth-map-loanamt-python.html" width="800" hei
```



Educational Loan

In [6]:
```
1 eda.mapdist_pur_loan_amt("educational")
```

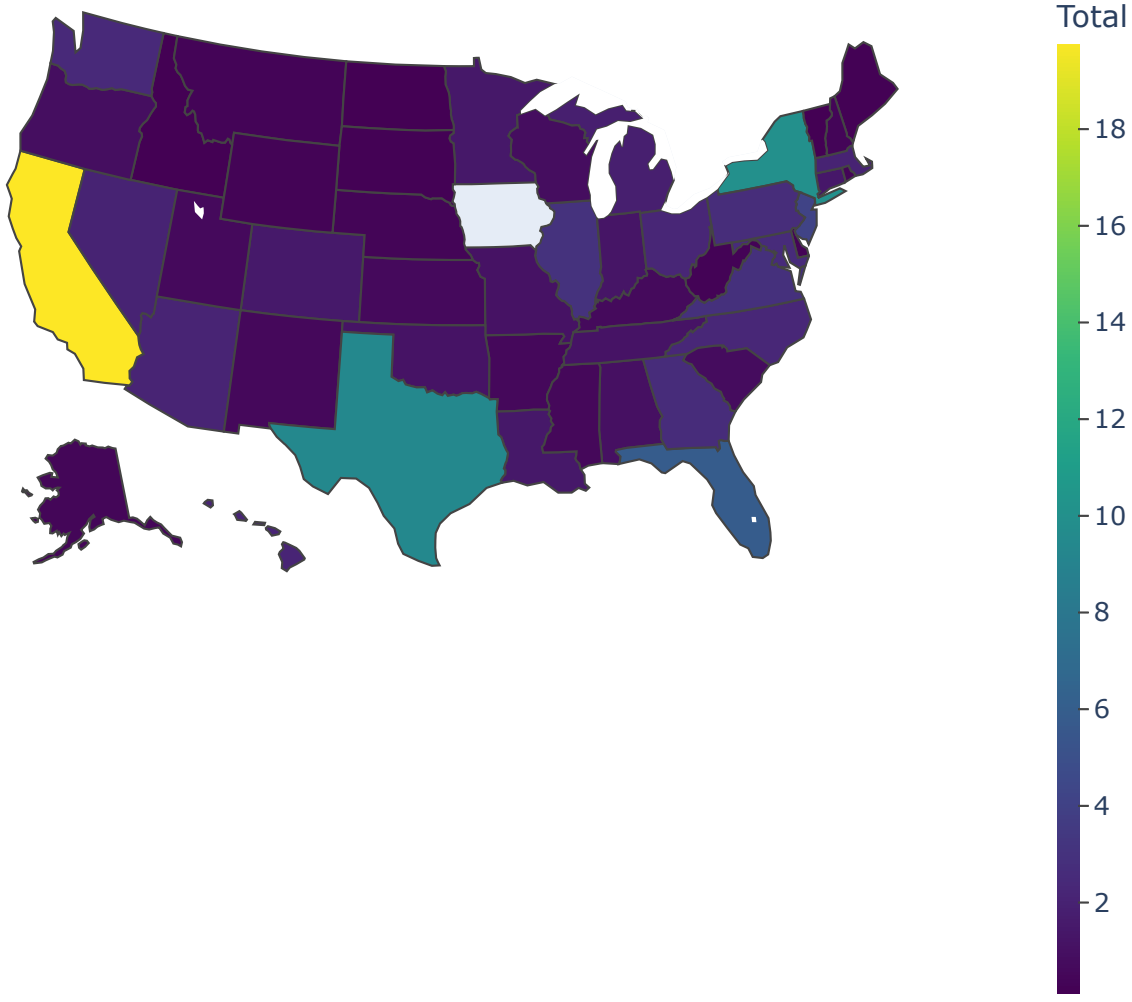
In [27]:
```
1 %%html
2 <iframe src="./educational-choropleth-map-loanamt-python.html" wid
```



Vacation Loan

In [7]:
```
eda.mapdist_pur_loan_amt("vacation")
```

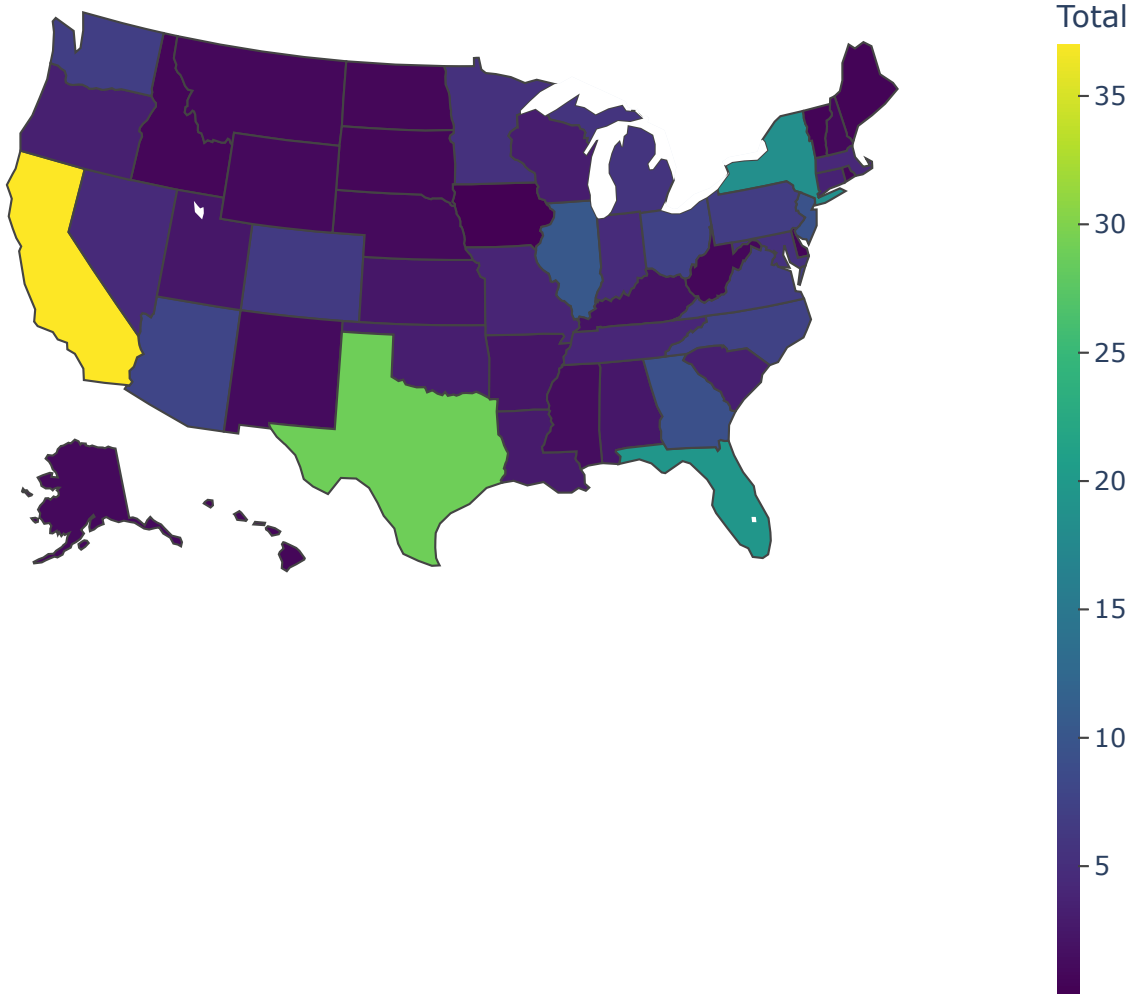
In [29]:
```
%%html
<iframe src="./vacation-choropleth-map-loanamt-python.html" width=
```



Medical Loan

In [8]:
```
eda.mapdist_pur_loan_amt("medical")
```

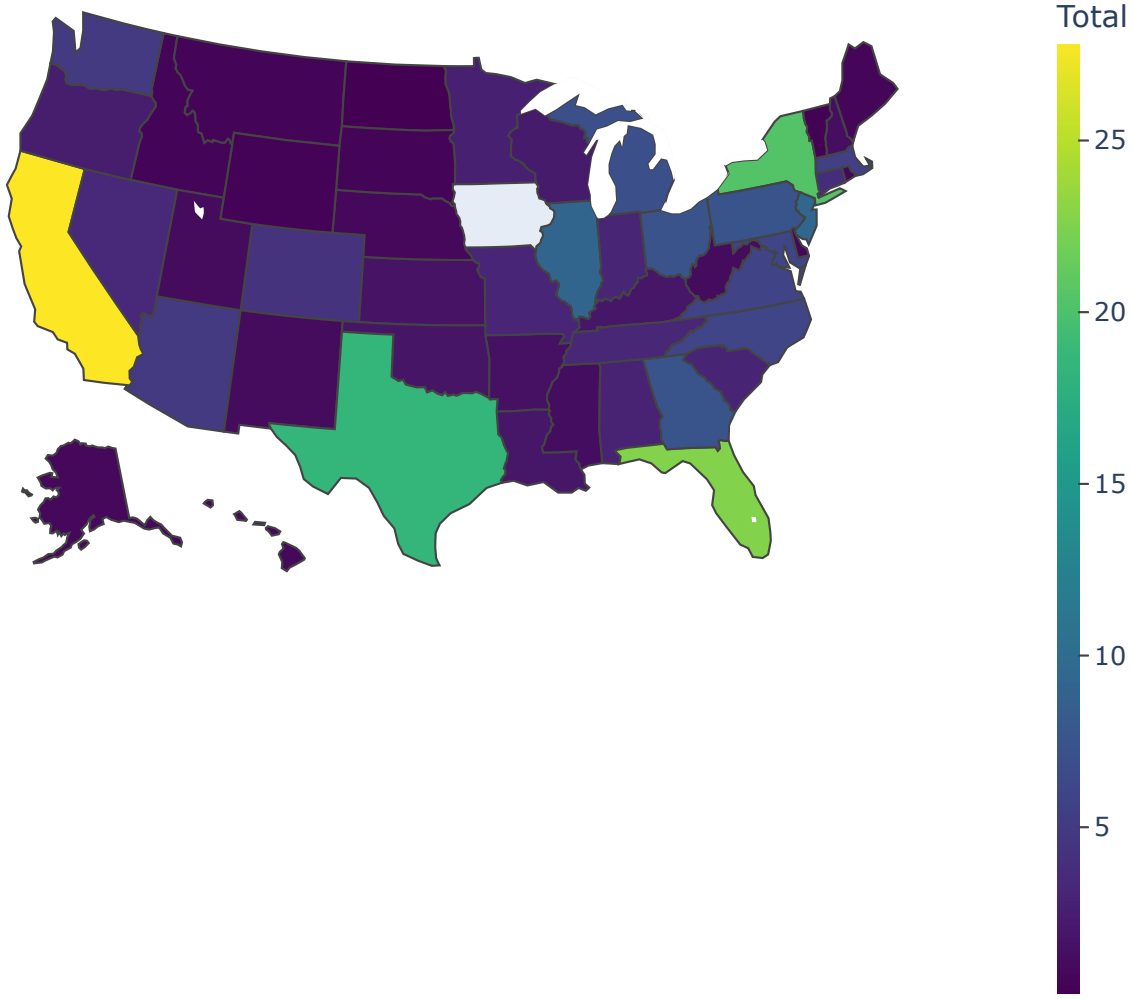
In [30]:
```
%%html
<iframe src="./medical-choropleth-map-loanamt-python.html" width="
```



House Loan

In [9]:
```
eda.mapdist_pur_loan_amt("house")
```

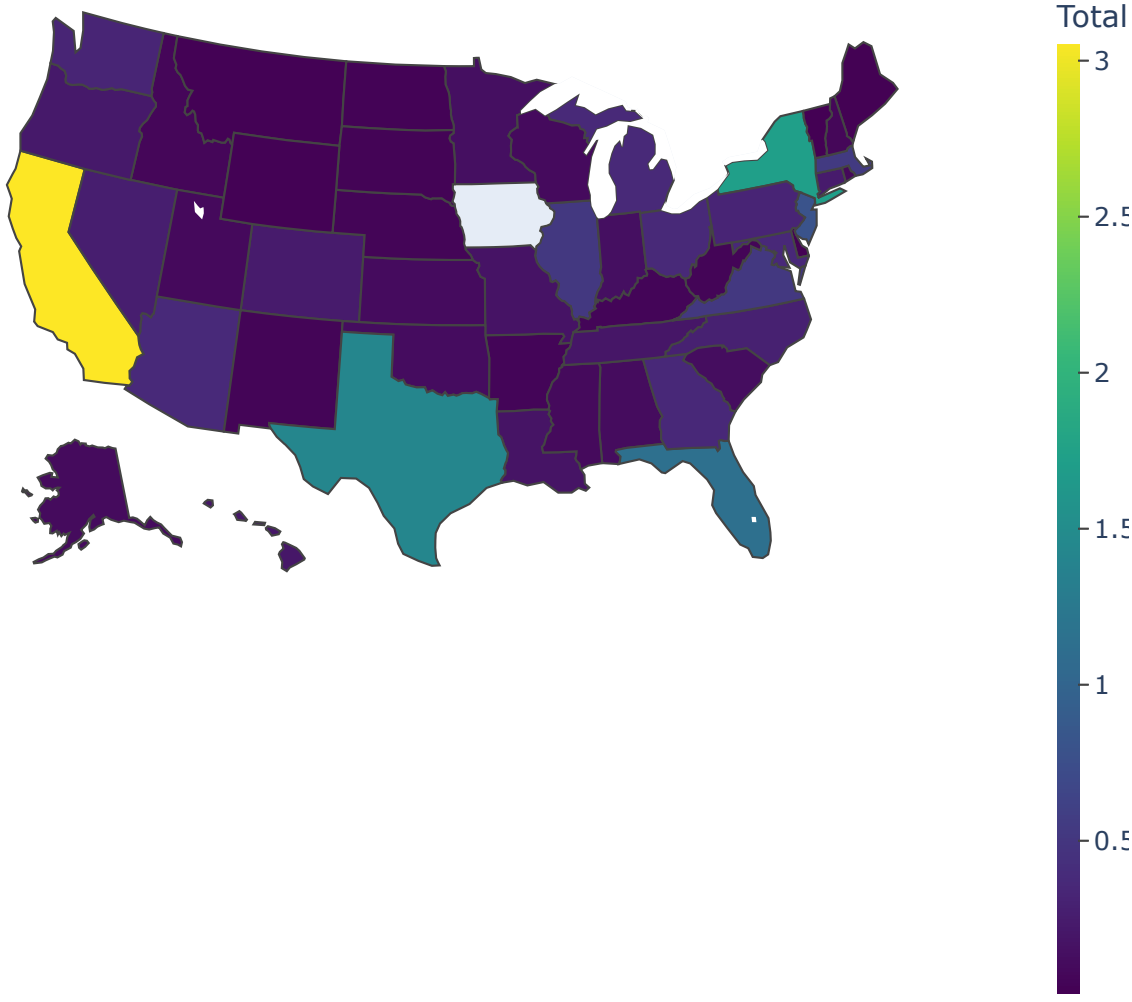
In [31]:
```
%%html
<iframe src="./house-choropleth-map-loanamt-python.html" width="80
```



Renewable Energy

In [10]:
```
eda.mapdist_pur_loan_amt("renewable_energy")
```

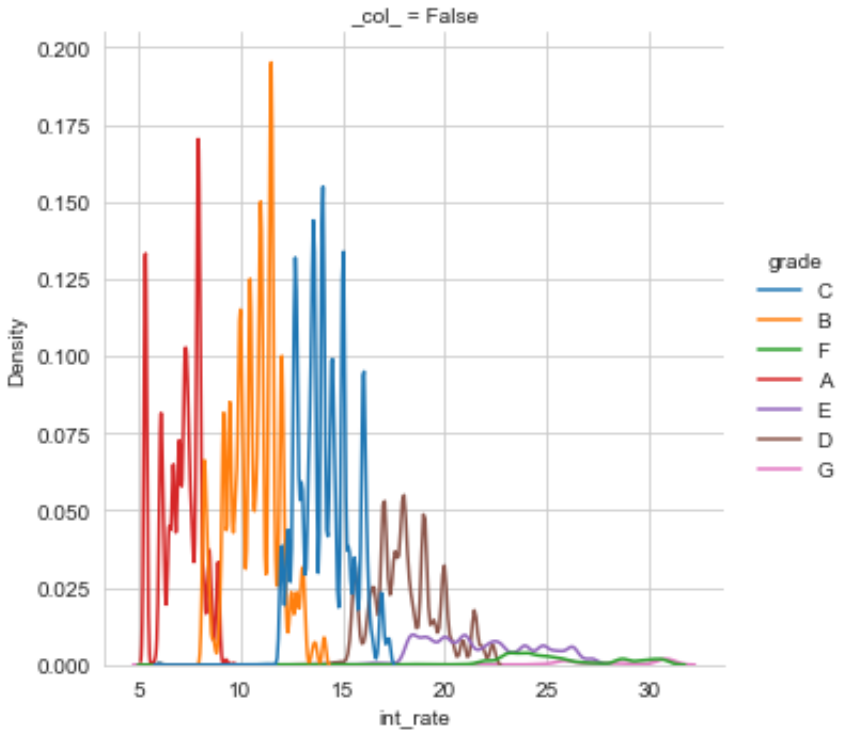
In [32]:
```
%%html
<iframe src="./renewable_energy-choropleth-map-loanamt-python.html
```



Loan Grade observations

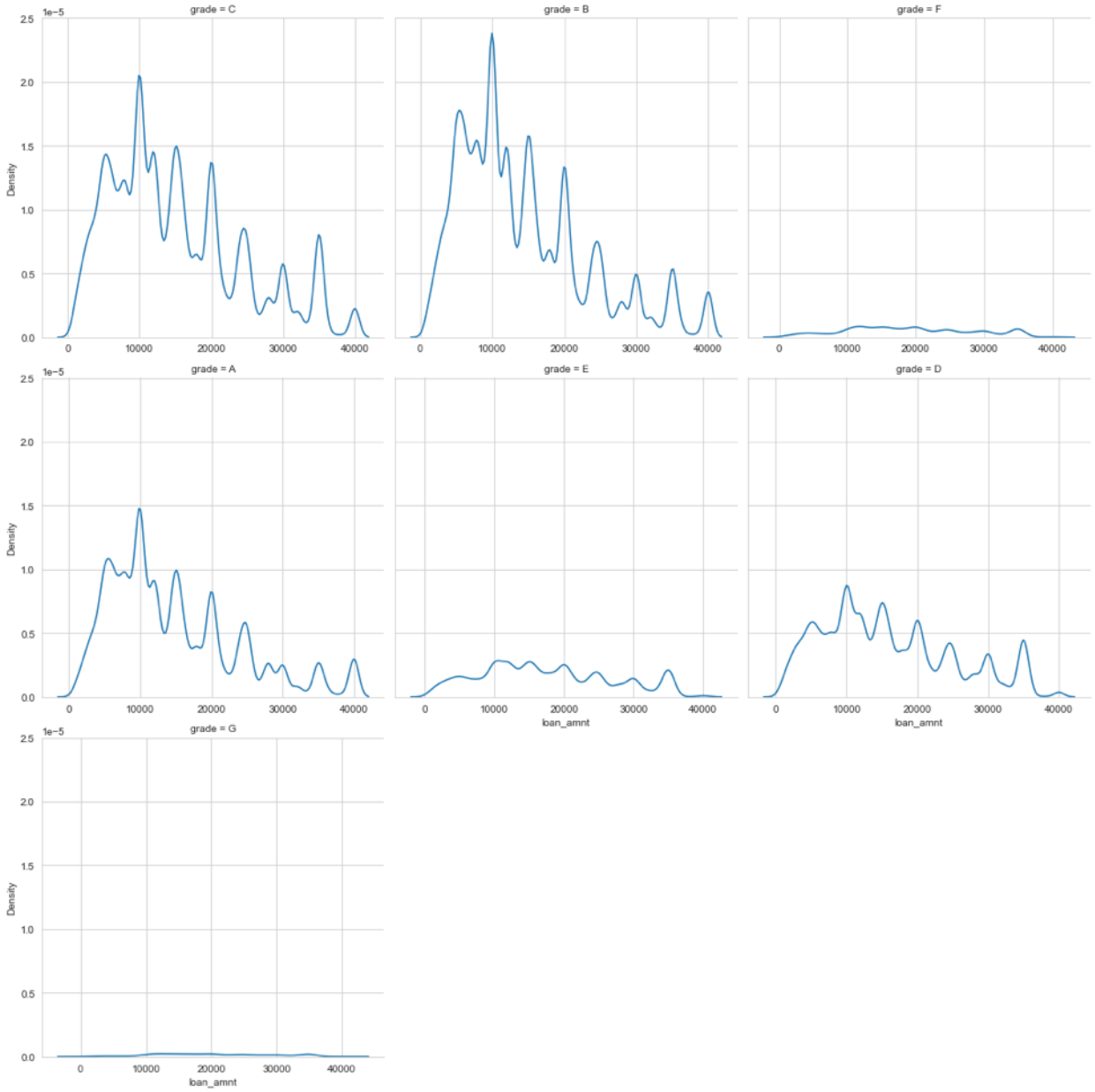
In [11]:
```
1 eda.simpledist(x='int_rate', hue ="grade")
```

```
<Figure size 1440x576 with 0 Axes>
```



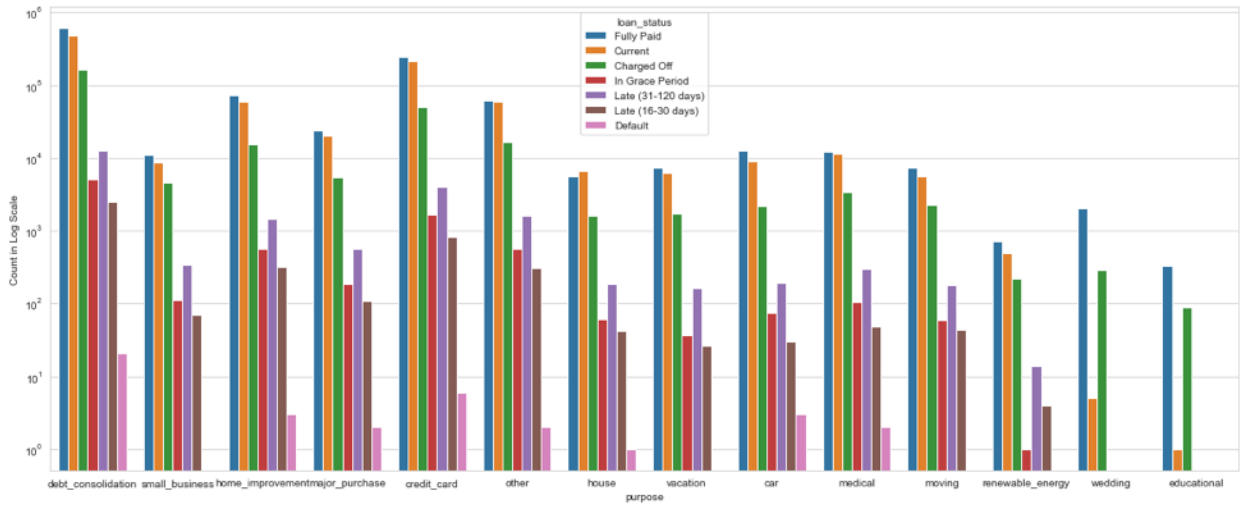
In [12]:
```
1 eda.comparedist(x='loan_amnt', hue =None,col="grade")
```

```
<Figure size 1440x576 with 0 Axes>
```

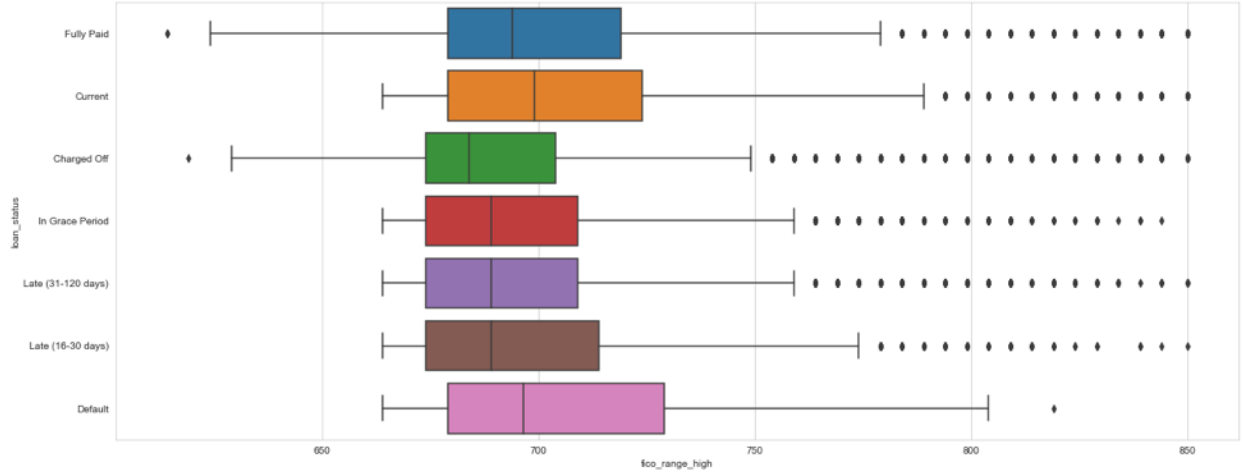


Loan Status/ FICO Observations

In [13]:
```
1 eda.logcountplot(x = "purpose",hue = "loan_status")
```

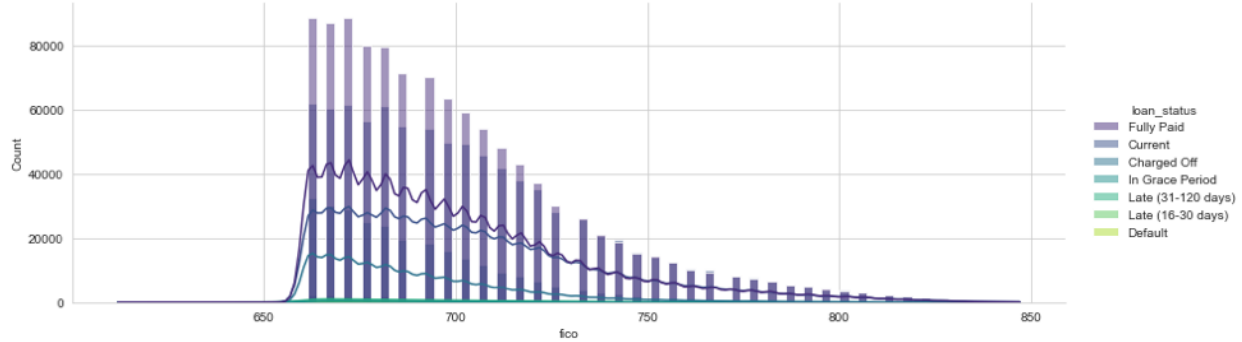


In [14]:
```
1 eda.boxplots(y='loan_status', x='fico_range_high')
```



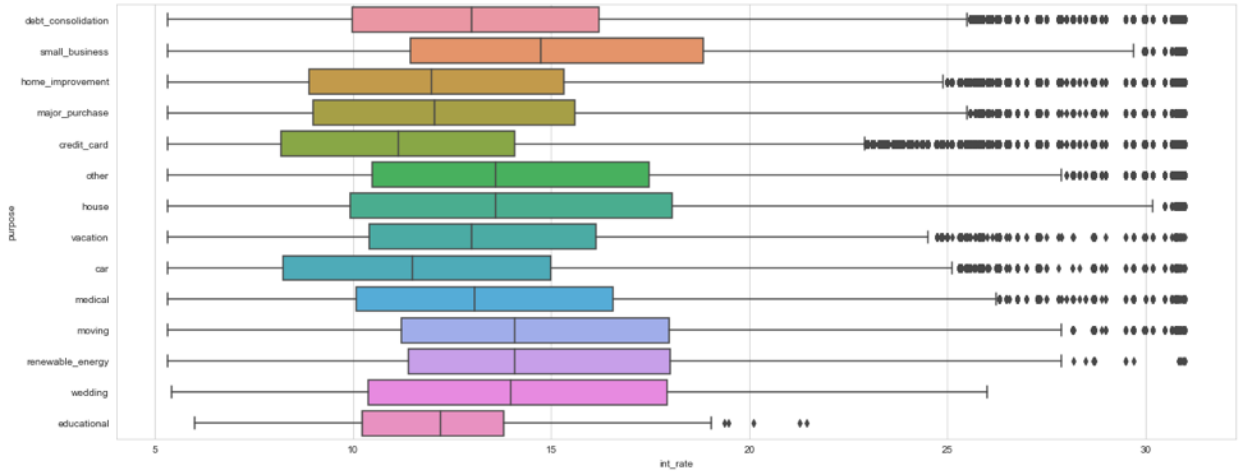
In [15]:
```
1 eda.fico_loan()
```

```
<Figure size 6000x2100 with 0 Axes>
```

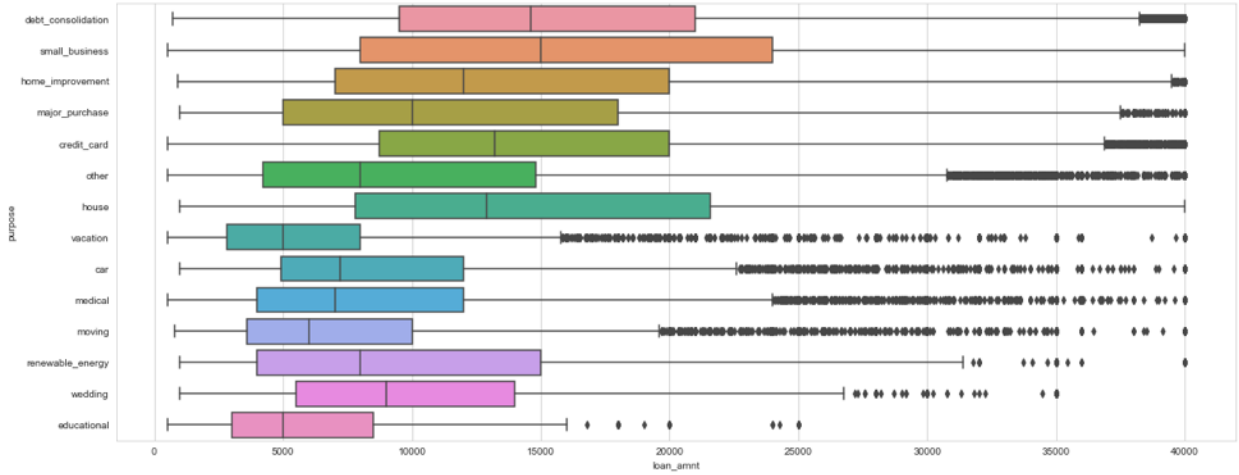


Loan Purpose Observations

In [16]:
```
1 eda.boxplots(x='int_rate',y='purpose')
```

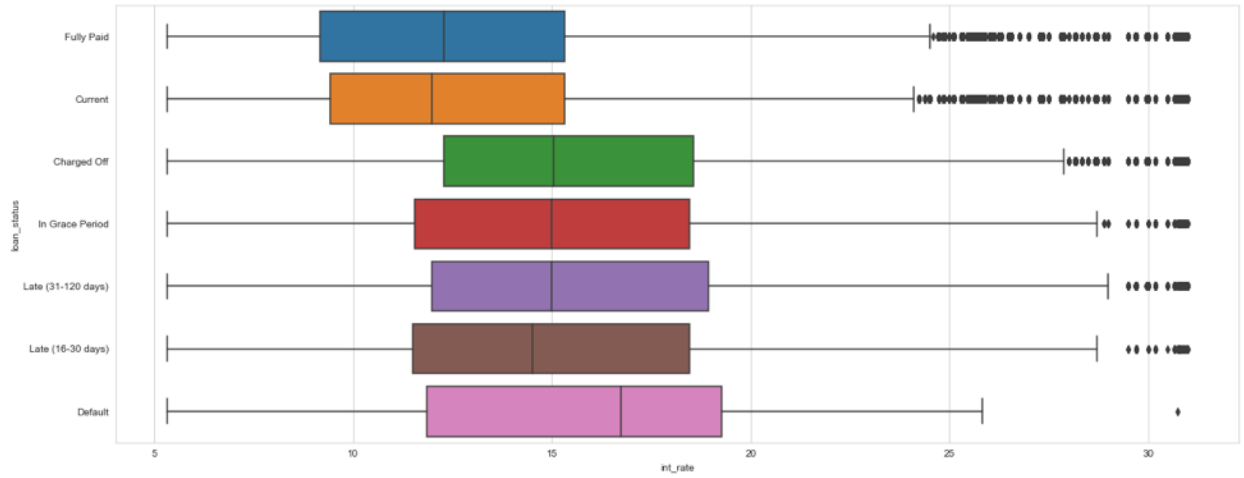


In [17]:
```
1 eda.boxplots(x='loan_amnt',y='purpose')
```

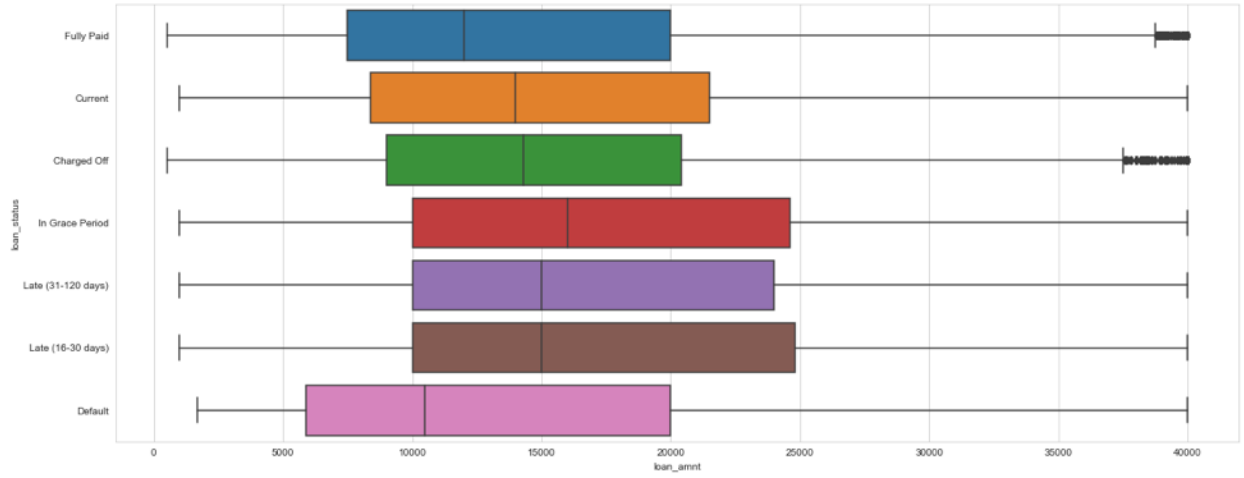


Loan Status Observations

In [18]:
```
1 eda.boxplots(x='int_rate',y='loan_status')
```

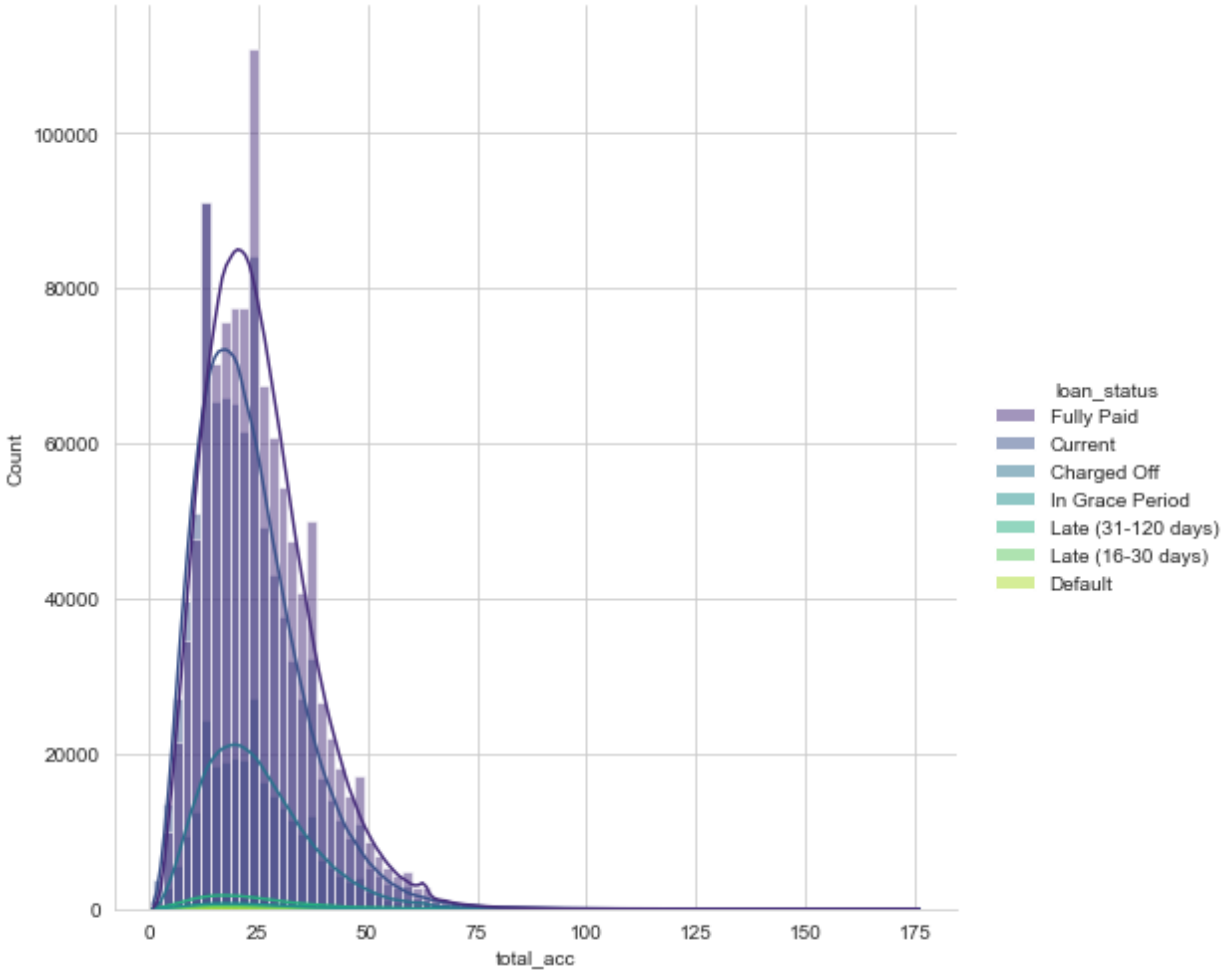


In [19]:
```
1 eda.boxplots(x='loan_amnt',y='loan_status')
```



Total no. of credit lines

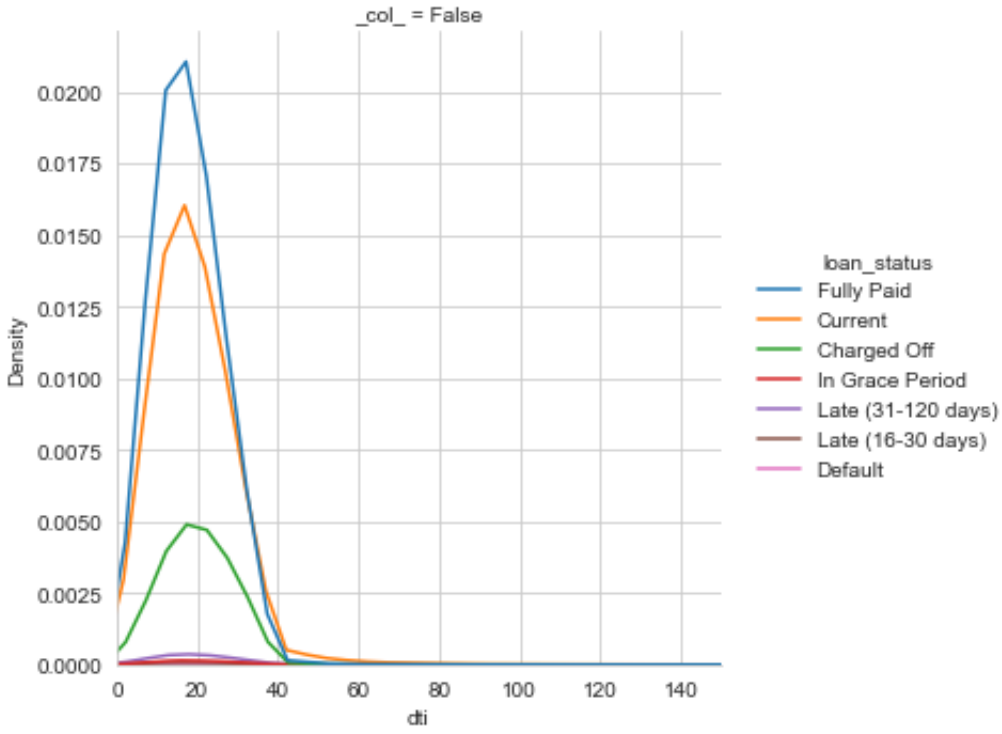
In [20]:
```
1 eda.comparehist(x='total_acc', hue ='loan_status')
```



Debt to Income Ratio

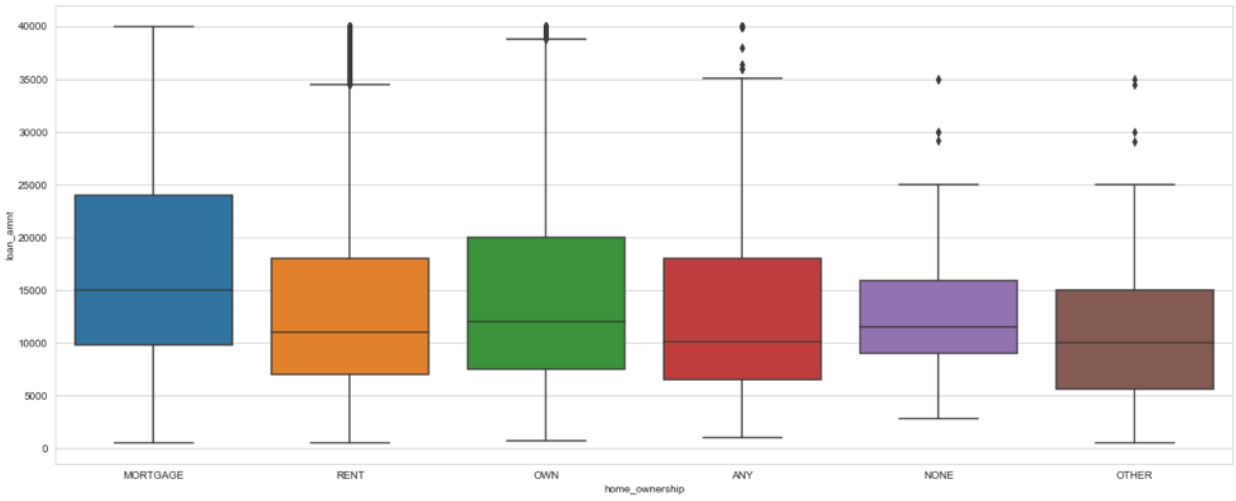
In [21]:
```
1 eda.dti_loanst(x='dti', hue ="loan_status")
```

```
<Figure size 1440x576 with 0 Axes>
```



Home Ownership vs Loan Amount

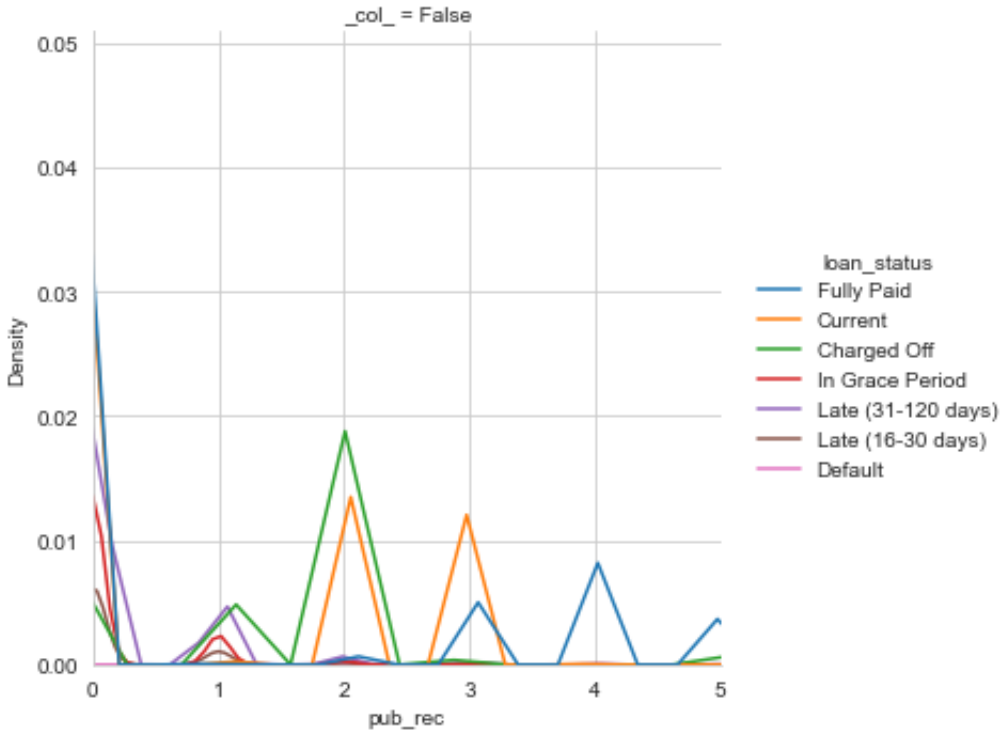
In [22]:
```
1 eda.boxplots(y='loan_amnt',x='home_ownership')
```



Public Derogatory Records vs Loan Status

In [23]:
```
1 eda.pubrec_simpledist(x='pub_rec',hue='loan_status')
```

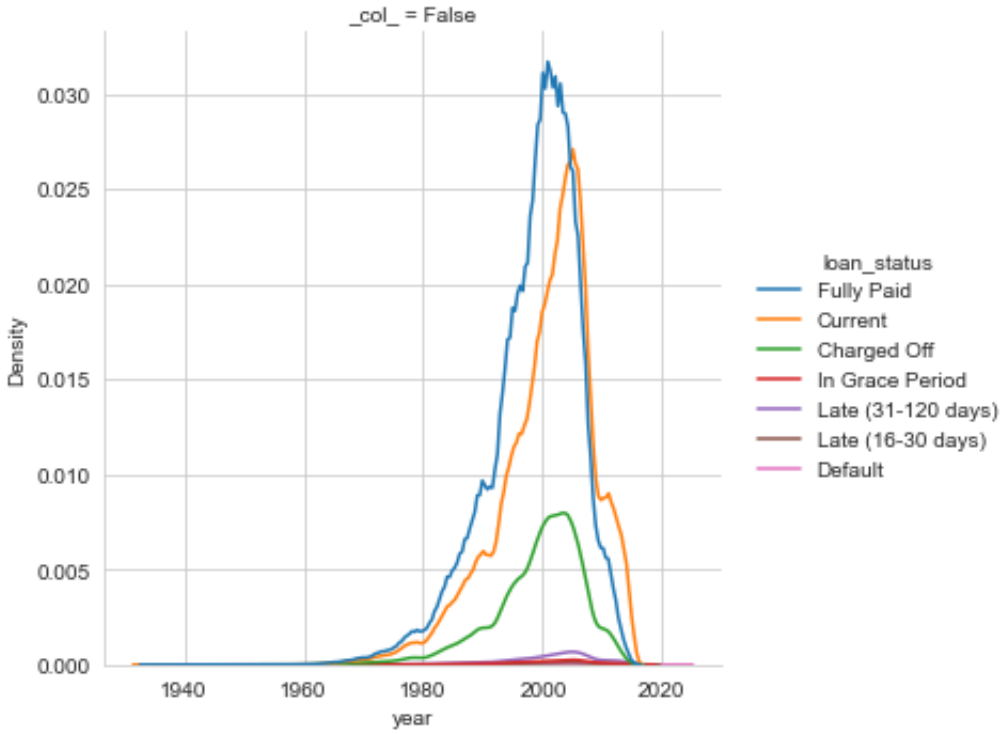
```
<Figure size 1440x576 with 0 Axes>
```



Earliest Credit Line Opening year vs Loan Status

In [24]:
```
1 eda.datedist(x='earliest_cr_line',hue='loan_status')
```

```
<Figure size 1440x576 with 0 Axes>
```



Observations:

1. Most of the loans were taken in the states of California, Texas, New York and Florida
2. People that paid the loan fully have higher incomes than that of people with loan status: current, charged off, late or default
3. Not just total loan, almost all choropleth ap plots show that the state of California has the highest loan amount borrowed. It may be the case of biased data sampling too.
4. Also, it is interesting to note that there are a few states without any educational loan taken. Again a possibility of data sampling bias.
5. Renewable energy loans are higher in the outer regions (CA, TX, NY etc) than the inner ones(NE,SD,WY).
6. As the grade increases alphabetically (starting from A), the interest rates go high.
7. These high interest rates lead to very few loans taken as evidenced by the low density of loan amounts borrowed in grades E and above.
8. When the mean of high and low of fico score is considered, the fico score of people who paid fully is higher than the ones that didnt pay yet.
9. Although the interest rates of "Medical" and "House" loans are similar, the loan amount borrowed for houses is far more.
10. Educational Loans have one of the lowest interest rates contrary to popular belief.
11. Although the interest rattes for Fully Paid and Curent Loans are similar, it appears that lower loan amounts have been paid back but the current, charged off and late ones are large amounts.
12. The total number of credit lines currently in the borrower's credit file (total_acc) shows that mean total_acc is higher for people who fully paid off the loan.
13. The debt-to-income (dti) ratio (total monthly debt payments divided by reported monthly income) is highest for people who fully paid the loan.
14. People on mortgage borrow significantly higher amount
15. People who paid off the loan fully have an earlier (mean) credit line opening date.