In [47]:
```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
%matplotlib inline
```

In [20]:
```
df=pd.read_csv('/Users/macintoshhd/Downloads/lendingclub_sample.csv
')
df['int_rate']=df['int_rate'].str.strip('%')
```

In [21]:
```
df.head()
```

Out[21]:

	id	member_id	loan_amnt	term	emp_title	emp_length	ho
0	55441634	59043359.0	18000.0	60 months	driver/wharehouseman	10+ years	M
1	38595688	41379463.0	18000.0	60 months	Supervisor	3 years	M
2	38455988	41249804.0	16000.0	36 months	Mail Clerk	9 years	O
3	40362356	43227157.0	4000.0	36 months	MANAGER INTERMODAL OPERATIONS	10+ years	RE
4	54207722	57748458.0	6000.0	36 months	Management	10+ years	M

5 rows × 29 columns

In [22]:
```
df.columns
```

Out[22]:
```
Index(['id', 'member_id', 'loan_amnt', 'term', 'emp_title', 'emp_l
ength',
       'home_ownership', 'annual_inc', 'verification_status', 'pur
pose',
       'zip_code', 'addr_state', 'dti', 'delinq_2yrs', 'earliest_c
r_line',
       'inq_last_6mths', 'mths_since_last_delinq',
       'mths_since_last_major_derog', 'mths_since_last_record', 'p
ub_rec',
       'open_acc', 'total_acc', 'acc_now_delinq', 'tot_coll_amt',
'revol_bal',
       'revol_util', 'total_credit_rv', 'tot_cur_bal', 'int_rate']
,
      dtype='object')
```

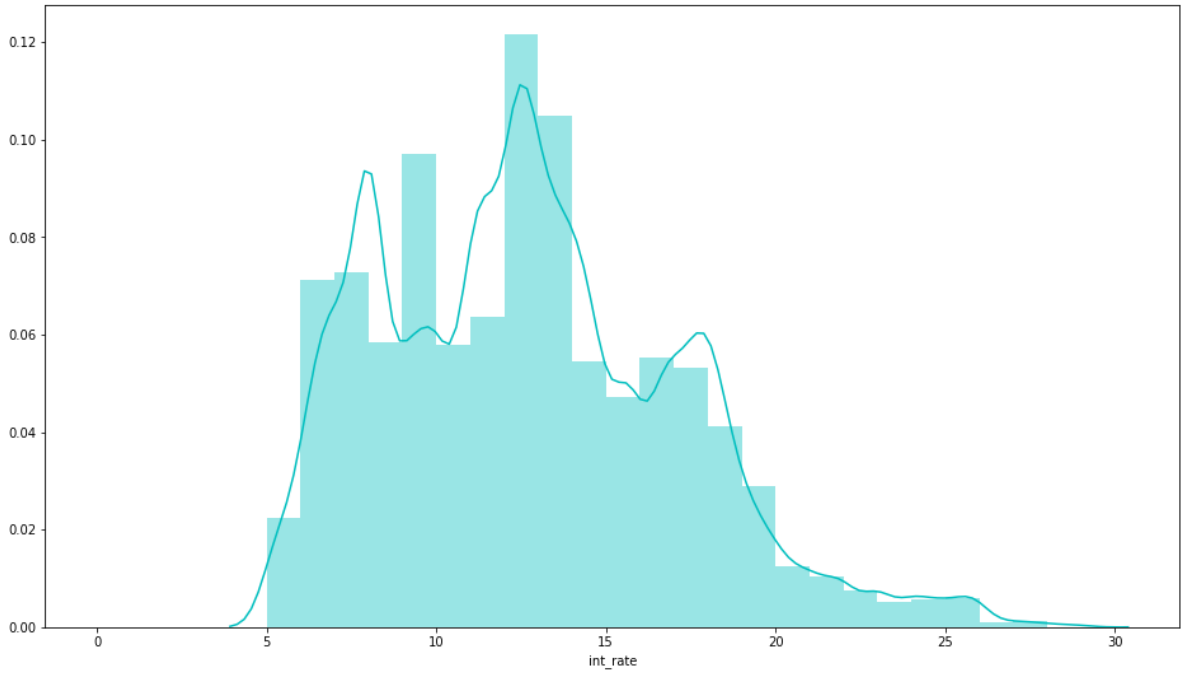
a

In [23]:
```
df['int_rate']=pd.to_numeric(df['int_rate'])
df['int_rate'].describe()
```

Out[23]:
```
count    99999.000000
mean        12.768029
std          4.392747
min          5.320000
25%          9.170000
50%         12.390000
75%         15.610000
max         28.990000
Name: int_rate, dtype: float64
```

In [24]:
```
plt.figure(figsize=(16,9))
sns.distplot(df['int_rate'],bins=np.arange(df['int_rate'].max()),color="c")
```

Out[24]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a165e2a58>
```



The highest interest rate is 28.99% in this dataset and the minimum value of interest rate is 5.32%, which means that the range of interest rate is 23.76%. The mean of interest rate is 12.76%.

From the plot, we can konw that the distribution of interest rate is roughly a right-tailed and most of the interest rate in sample dataset fall between 5%-20%. The most frequent interest rate is about 13%, 14% following. The number of cases decrease when interest rate increases after 17%.

b

In [25]:
```
df.term.unique()
```

Out[25]:
```
array([' 60 months', ' 36 months'], dtype=object)
```

In [26]:
```
df1=df[df['term']==' 60 months']
df2=df[df['term']==' 36 months']
```

In [27]:
```
# 60-month int_rate
df1['int_rate'].describe()
```

Out[27]:
```
count    33108.000000
mean        15.490021
std          4.347654
min          6.000000
25%         12.290000
50%         14.650000
75%         18.250000
max         28.990000
Name: int_rate, dtype: float64
```

In [28]:
```
# 36-month int_rate
df2['int_rate'].describe()
```

Out[28]:
```
count    66891.000000
mean        11.420767
std          3.742880
min          5.320000
25%          8.180000
50%         11.530000
75%         13.990000
max         28.990000
Name: int_rate, dtype: float64
```

In [29]:
```
# 60-month loan_amnt
df1['loan_amnt'].describe()
```

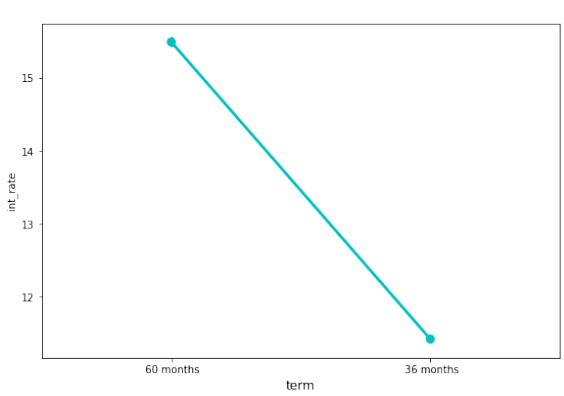
Out[29]:
```
count    33108.000000
mean     20211.971427
std       7320.135867
min      10000.000000
25%      14400.000000
50%      19375.000000
75%      25000.000000
max      35000.000000
Name: loan_amnt, dtype: float64
```

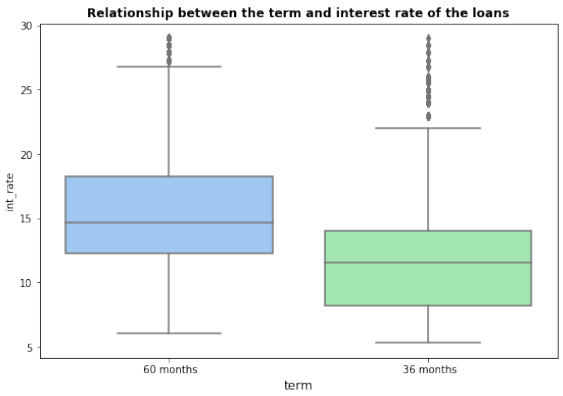
In [30]:
```
# 36-month loan_amnt
df2['loan_amnt'].describe()
```

Out[30]:
```
count    66891.000000
mean     12831.162264
std       8030.618181
min       1000.000000
25%       7000.000000
50%      10000.000000
75%      17000.000000
max      35000.000000
Name: loan_amnt, dtype: float64
```

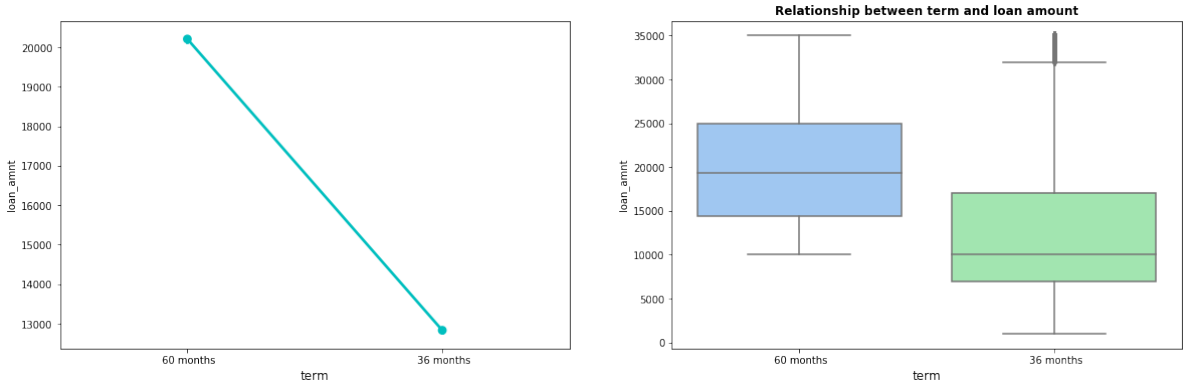
In [31]:
```
fig, ax = plt.subplots(1, 2, sharex=True, figsize=(20,6))
plt.title('Relationship between the term and interest rate of the l
oans',fontweight='bold', fontsize=12);
sns.pointplot(x='term', y='int_rate', data=df,ax=ax[0],color="c")
sns.boxplot(x='term', y='int_rate', data=df, palette=sns.color_pale
tte('pastel'),ax=ax[1])
ax[0].set_xlabel('term', fontsize=12)
ax[1].set_xlabel('term', fontsize=12)
```

Out[31]:
```
Text(0.5,0,'term')
```





In [32]:
```
fig, ax = plt.subplots(1, 2, sharex=True, figsize=(20,6))
sns.pointplot(x='term', y='loan_amnt', data=df,ax=ax[0],color="c")
sns.boxplot(x='term', y='loan_amnt', data=df, palette=sns.color_pal
ette('pastel'))
ax[0].set_xlabel('term', fontsize=12)
ax[1].set_xlabel('term', fontsize=12)
plt.title('Relationship between term and loan amount', fontweight='
bold', fontsize=12);
```



From the plot, we can know that comparing with 60-month term, interest rate and loan amount of 36-month term are lower Generally, the shorter the term, the smaller of the loan amount. Also, the shorter the term, the less risk of loan default and thus lower interest rate.

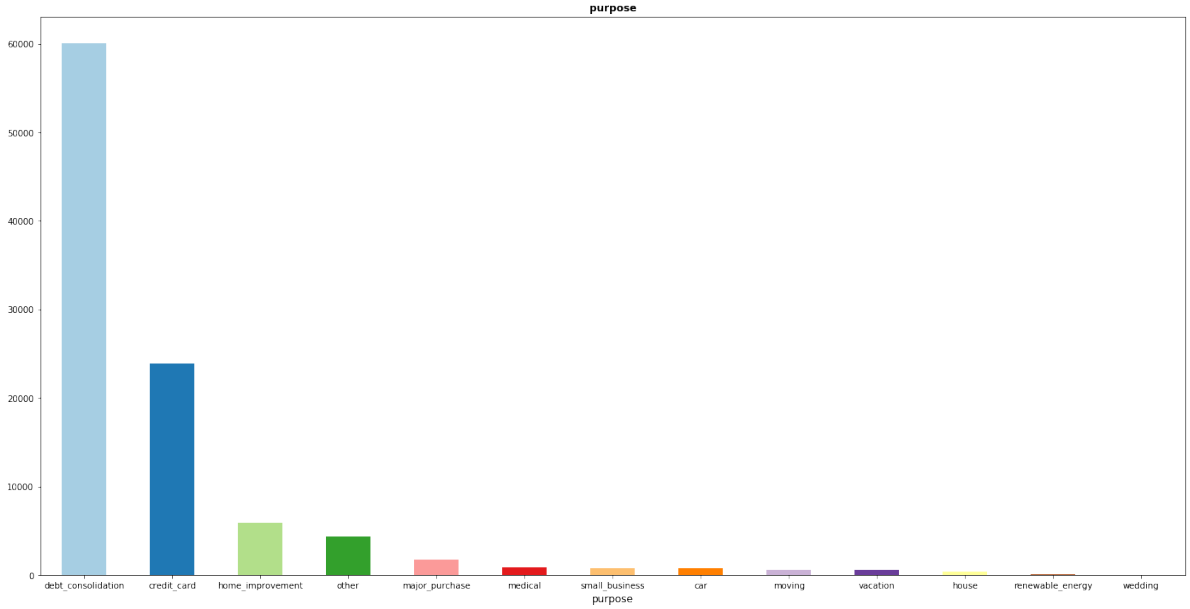
C

In [33]:
```
df['purpose'].value_counts()
```

Out[33]:
```
debt_consolidation    60108
credit_card           23878
home_improvement       5905
other                  4343
major_purchase         1752
medical                 842
small_business          792
car                     787
moving                  608
vacation                581
house                   348
renewable_energy         54
wedding                   1
Name: purpose, dtype: int64
```

In [34]:
```
df['purpose'].value_counts().plot(kind='bar', figsize=(24,12), rot=
0, color=[plt.cm.Paired(np.arange(len(df['purpose'].unique())))])
plt.title('purpose', fontweight='bold', fontsize=12)
plt.xlabel('purpose', fontsize=12)
```

Out[34]: Text(0.5,0,'purpose')



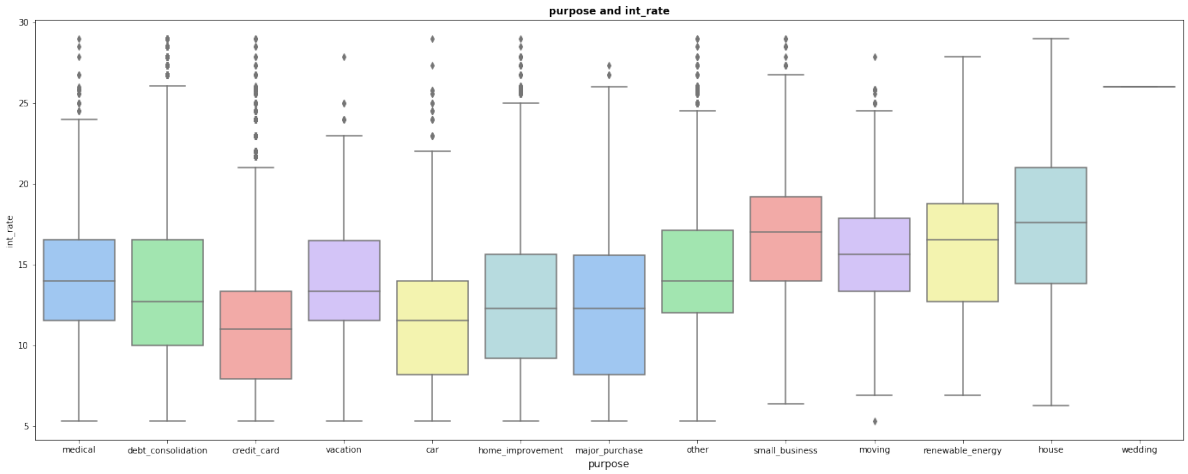
Most of the loans are used for debt_consolidation purpose and for credit_card following. Only one borrower applied loan for their wedding.

In [35]:
```
df.groupby(['purpose'])['int_rate'].mean().sort_values(ascending=Fa
lse)
```

Out[35]:
```
purpose
wedding               25.990000
house                 17.547069
small_business        17.092235
renewable_energy      16.243148
moving                15.708026
other                 14.579417
medical               13.993349
vacation              13.584096
debt_consolidation    13.183126
home_improvement      12.760588
major_purchase        12.567340
car                   11.888806
credit_card           11.079774
Name: int_rate, dtype: float64
```

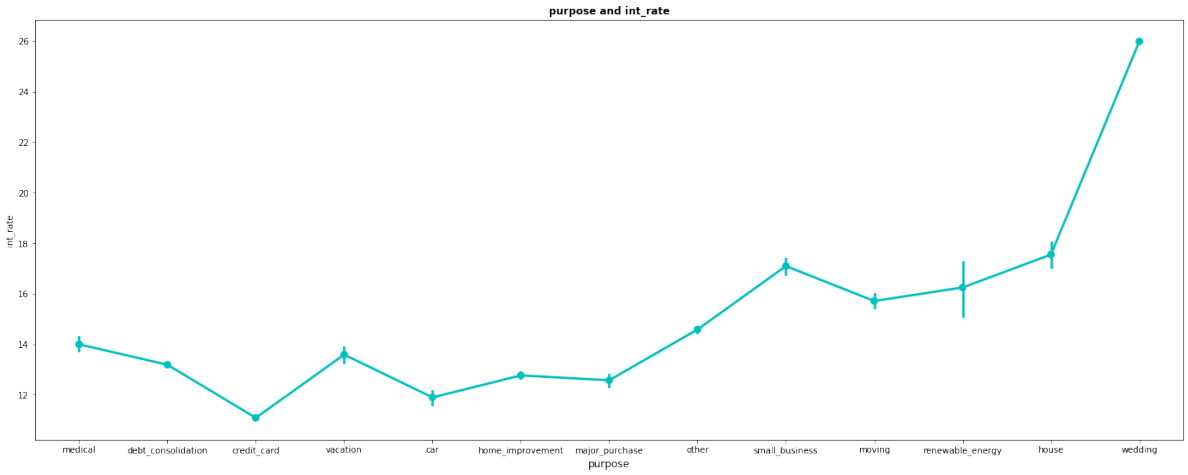
In [36]:
```
plt.figure(figsize=(24,9))
sns.boxplot(x='purpose', y='int_rate', data=df, palette=sns.color_palette('pastel'))
plt.title('purpose and int_rate', fontweight='bold', fontsize=12)
plt.xlabel('purpose', fontsize=12)
```

Out[36]: Text(0.5,0,'purpose')



In [37]:
```
plt.figure(figsize=(24,9))
sns.pointplot(x='purpose', y='int_rate', data=df,color="c")
plt.title('purpose and int_rate', fontweight='bold', fontsize=12)
plt.xlabel('purpose', fontsize=12)
```

Out[37]: Text(0.5,0,'purpose')



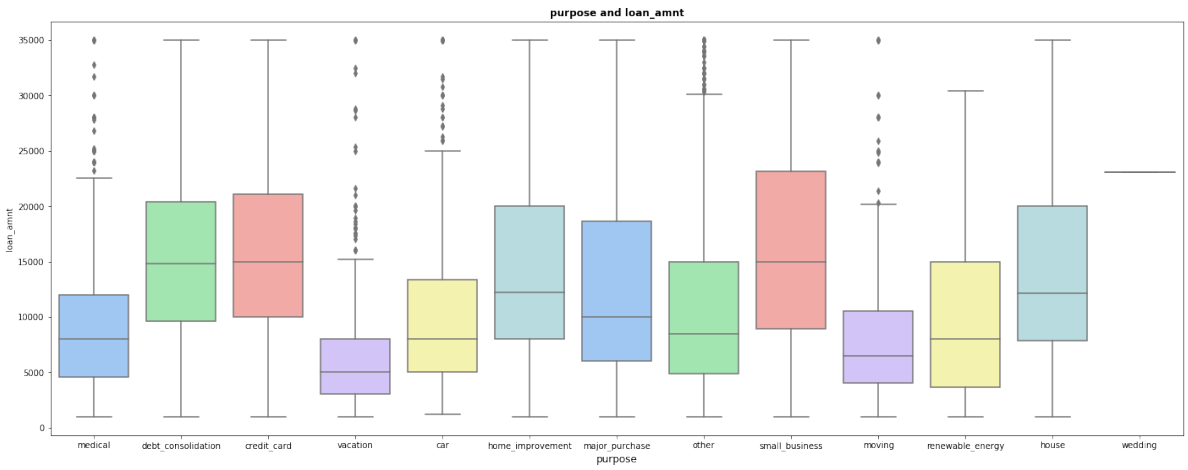
The mean interest rate for wedding is highest among other purposes and housing following. The reason that wedding purpose has highest interest rate is that only one case in dataset is for wedding purpose, so the case have huge influence for this purpose category. The reason of housing is that the number of cases of hosing purpose is small in dataset and the housing market is not optimistic in these years, so lending money for housing purpose is more risky. The interest rate for credit_card purpose is that a lot of cases in dataset is for credit card purpose and thus the mean will be smaller, and that the lowest because people who have credit card have been evaluated by credit card organization and probably have more complete credit records. Therefore, the interest rate for credit card is lowest. The range of interest rate for house is the largest because house value may vary greatly.

In [38]:
```
df.groupby(['purpose'])['loan_amnt'].mean().sort_values(ascending=False)
```

Out[38]:
```
purpose
wedding                23100.000000
small_business         16247.348485
credit_card            16065.720538
debt_consolidation     15711.832868
home_improvement       14840.469941
house                  14738.290230
major_purchase         13131.207192
other                  10461.892701
renewable_energy       10237.500000
car                    10229.606099
medical                 9287.767221
moving                  8379.152961
vacation                6592.039587
Name: loan_amnt, dtype: float64
```

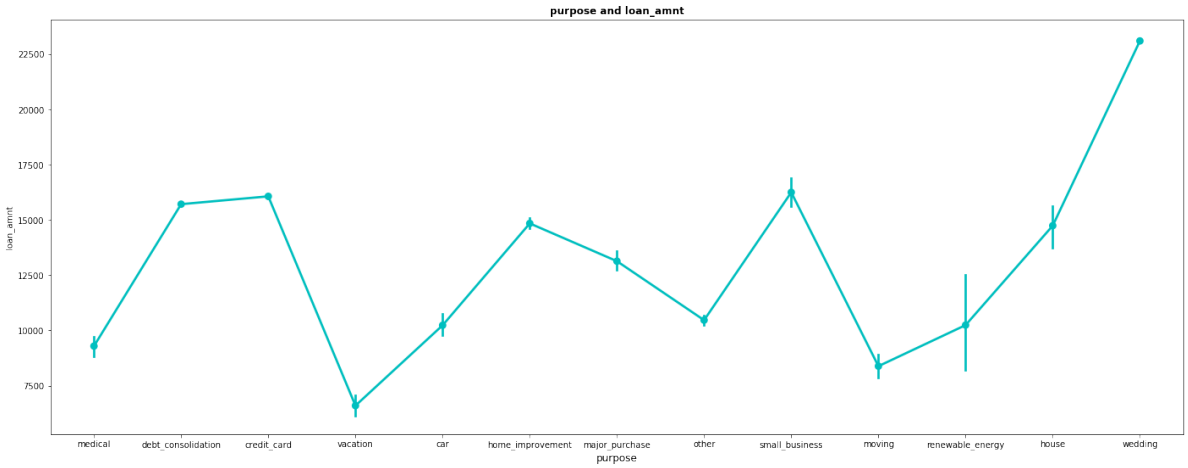
In [39]:
```
plt.figure(figsize=(24,9))
sns.boxplot(x='purpose', y='loan_amnt', data=df, palette=sns.color_palette('pastel'))
plt.title('purpose and loan_amnt', fontweight='bold', fontsize=12)
plt.xlabel('purpose', fontsize=12)
```

Out[39]:
```
Text(0.5,0,'purpose')
```



In [40]:
```
plt.figure(figsize=(24,9))
sns.pointplot(x='purpose', y='loan_amnt', data=df,color="c")
plt.title('purpose and loan_amnt', fontweight='bold', fontsize=12)
plt.xlabel('purpose', fontsize=12)
```

Out[40]: Text(0.5,0,'purpose')



The mean loan amount for wedding purpose is highest among the other purpose and small business following. The reason that wedding purpose has highest loan amount is that only one case in dataset is for wedding purpose so the case have huge influence. The large loan amount for small business purpose is that fund requirement of running small businss is large. The large loan amount for credit_card purpose is that people who have credit card have been evaluated by credit card organization and probably have more complete credit records and thus being grant higher loan amount. The mean loan amount for vocation is smallest because fund requirement for vocation is not large.

d

In [41]:
```
import datetime as dt
df['earliest_cr_line']=pd.to_datetime(df['earliest_cr_line'])
data_dt= dt.datetime(year=2015,day=30,month=9)
df['length']=(data_dt-df['earliest_cr_line'])
df['length']=df['length'].dt.days.astype(int)/365.25
df.insert(30,'credit_history',pd.cut(df['length'],np.arange(0,70,5)
))
df.head()
```

Out[41]:

	id	member_id	loan_amnt	term	emp_title	emp_length	ho
0	55441634	59043359.0	18000.0	60 months	driver/wharehouseman	10+ years	M(
1	38595688	41379463.0	18000.0	60 months	Supervisor	3 years	M(
2	38455988	41249804.0	16000.0	36 months	Mail Clerk	9 years	O\
3	40362356	43227157.0	4000.0	36 months	MANAGER INTERMODAL OPERATIONS	10+ years	RE
4	54207722	57748458.0	6000.0	36 months	Management	10+ years	M(

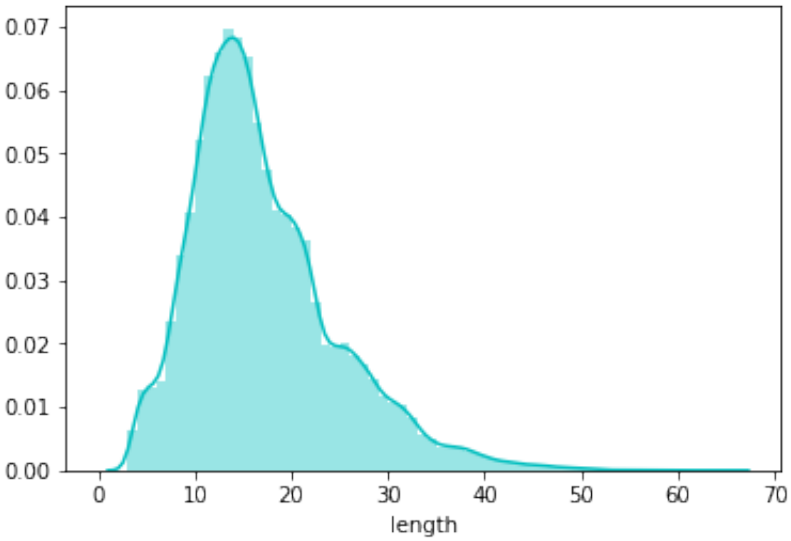
5 rows × 31 columns

In [42]:
```
df['length'].describe()
```

Out[42]:
```
count    99999.000000
mean        17.134729
std          7.587571
min          3.162218
25%         11.912389
50%         15.581109
75%         21.078713
max         65.163587
Name: length, dtype: float64
```

In [43]:
```
sns.distplot(df['length'],bins=np.arange(df['length'].max()),color="c")
```

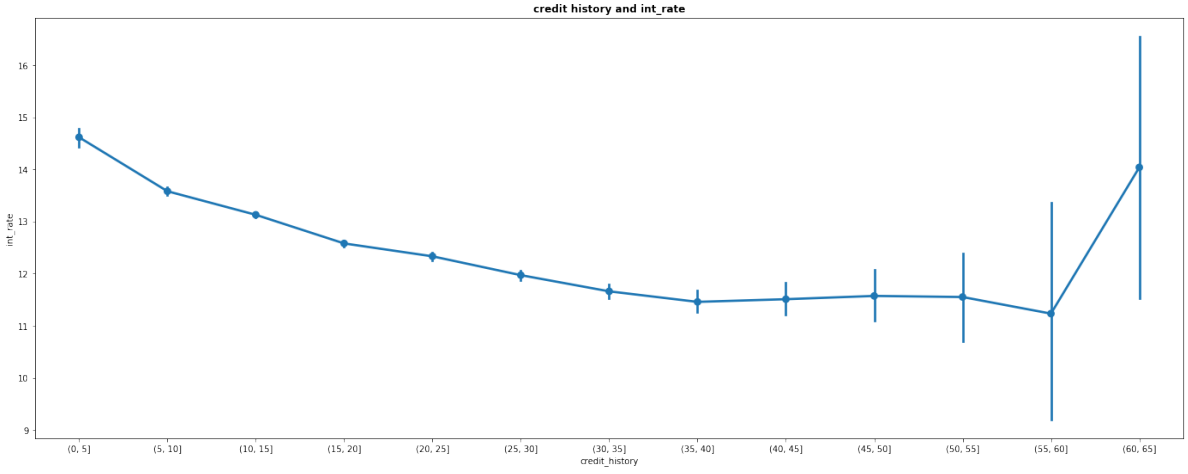
Out[43]: <matplotlib.axes._subplots.AxesSubplot at 0x1a163f6198>



In the sample data set, the max credit history is 65.16 years and the minimum is 3.16 years, so the range is 62 year. The mean of credit history is 17.13 years. Most of the cases in the dataset fall between 10-20 year.

In [44]:
```
plt.figure(figsize=(24,9))
sns.pointplot(x='credit_history', y=df['int_rate'], data=df)
plt.title('credit history and int_rate', fontweight='bold', fontsize=12)
```

Out[44]: Text(0.5,1,'credit history and int_rate')



From the graph, we can know that the interest rate decreases along with increasing credit history, but decrease after credit interval [55,60]. Normally, the longer your credit history, the more accurate lender can be in determining the level of risk it takes. Thus, long and good credit history enables borrower a lower interest rate. However, after credit history longer than 60 years, borrowers are old and may be retired. Therefore, their repayment ability decrease. Under such circumstance, lenders tend to increase interest rate to pretect their interest.

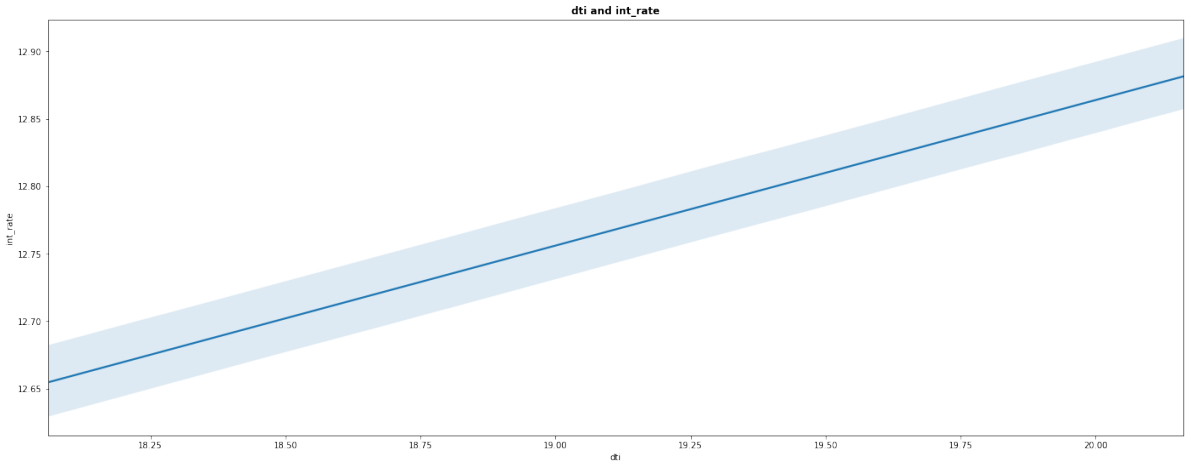
e

In [45]:
```
df['dti'].describe()
```

Out[45]:
```
count    99999.000000
mean        19.111652
std          8.623311
min          0.000000
25%         12.610000
50%         18.570000
75%         25.270000
max         39.990000
Name: dti, dtype: float64
```

In [58]:
```
plt.figure(figsize=(24,9))
sns.regplot(x='dti', y=df['int_rate'], data=df,scatter=False)
plt.title('dti and int_rate', fontweight='bold', fontsize=12)
```

Out[58]:
```
Text(0.5,1,'dti and int_rate')
```



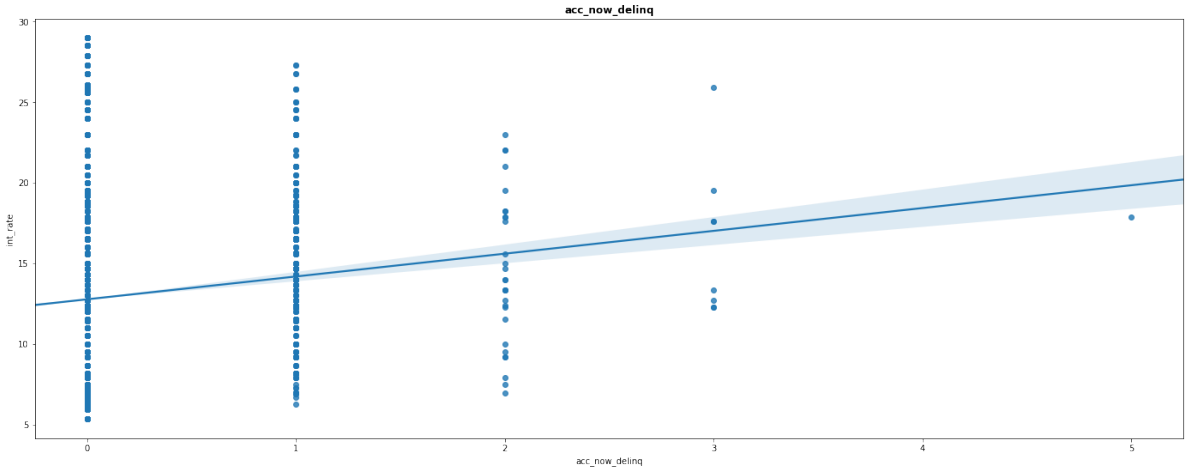
dti is selected as a debt variable to analyze the relationship with int_rate. This positive relationship of the regplot indicates that the higher the dti ratio, the higher the interest rate. This is becasue the high dti ratio means that borrowers spend a significant proportion of their income on debt payment, which represents unhealthy financial situation of the borrowers, and that the default risk of borrower is high. Thus, the int_rate increases.

In [55]:
```
df['acc_now_delinq']
```

Out[55]:
```
array([ 0.,  1.,  3.,  2.,  5.])
```

In [59]:
```
plt.figure(figsize=(24,9))
sns.regplot(x='acc_now_delinq', y=df['int_rate'], data=df)
plt.title('acc_now_delinq', fontweight='bold', fontsize=12)
```

Out[59]:
```
Text(0.5,1,'acc_now_delinq')
```



From the plot, we can know that the number of accounts on which the borrower is now delinquent, the higher the interest rate. When people have more delinquent account, they may be more likely to delinquent on the future loan repayment and thus more risky. Correspondingly, the interest rate for such borrowers will increase.