LENDING CLUB CASE STUDY

By Rishabh Vij

Case Study Objective

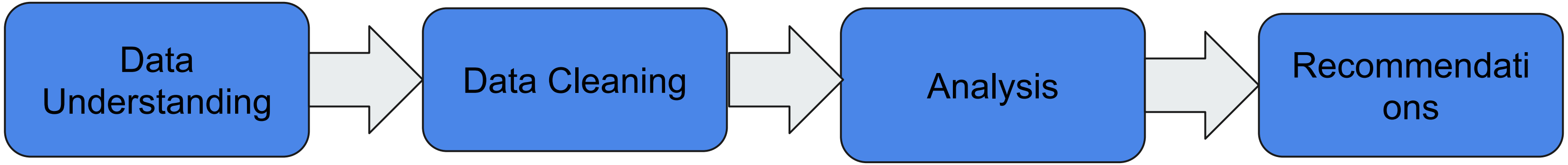
Background

Lending Club company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Problem Statement

Lending Club wants to identify the risky loan applicants, so that such loans can be reduced thereby cutting down the amount of credit loss. Our aim is to find the driving factors and features behind loan getting defaulted so that appropriate measures can be taken to mitigate this.

Solution Strategy



Understanding the Data

Data Inputs

loan.csv - csv file of shape - 39717, 111. This is our primary dataset

Data_Dictionary.xlsx - xlsx file of shape - 115, 2. Explains the Loan Dataset features

Our Main loan dataset, initially has features of following data type:

float64(74), int64(13), object(24)

Data Cleaning

The following steps were taken to clean the data:

- Missing Values were treated
- Outliers detected and handled
- Values Transformation (fixing and readying values for analysis)
- Features that aren't relevant to us were dropped - the likes of metadata, behavioural features, identifiers, etc.
- Derived Metrics to augment new feature columns like issue year and bucketed income & loan amount.

Post Cleaning the data out final dataset shape is - 38448, 22

Here's the final features which we use for our analysis -

```
[108]: df.columns
       Last executed at 2022-05-11 21:57:35 in 64ms

[108]: Index(['funded_amnt_inv', 'term_months', 'int_rate_pct', 'installment',
              'grade', 'sub_grade', 'emp_title', 'emp_length', 'home_ownership',
              'annual_inc', 'verification_status', 'issue_d', 'loan_status',
              'purpose', 'title', 'zip_code', 'addr_state', 'dti',
              'anual_income_bucketed', 'funded_amnt_inv_bucketed', 'issue_year',
              'interest_rate_bucketed'],
             dtype='object')
```

Data Analysis (Univariate)

Exploring numerical features with summary metrics

```
[51]: df.select_dtypes([float, int,]).describe().apply(lambda s: s.apply('{0:.5f}'.format)) ## exploring numeric features
Last executed at 2022-05-11 20:26:46 in 84ms
```

[51]:

	loan_amnt	funded_amnt	funded_amnt_inv	term_months	int_rate_pct	installment	annual_inc	dti	issue_year
count	38577.00000	38577.00000	38577.00000	38577.00000	38577.00000	38577.00000	38577.00000	38577.00000	38577.00000
mean	11047.02543	10784.05851	10222.48112	41.89844	11.93222	322.46632	68777.97368	13.27273	2010.30907
std	7348.44165	7090.30603	7022.72064	10.33314	3.69133	208.63921	64218.68180	6.67304	0.88266
min	500.00000	500.00000	0.00000	36.00000	5.42000	15.69000	4000.00000	0.00000	2007.00000
25%	5300.00000	5200.00000	5000.00000	36.00000	8.94000	165.74000	40000.00000	8.13000	2010.00000
50%	9600.00000	9550.00000	8733.44000	36.00000	11.71000	277.86000	58868.00000	13.37000	2011.00000
75%	15000.00000	15000.00000	14000.00000	36.00000	14.38000	425.55000	82000.00000	18.56000	2011.00000
max	35000.00000	35000.00000	35000.00000	60.00000	24.40000	1305.19000	6000000.00000	29.99000	2011.00000

Data Analysis (Univariate)

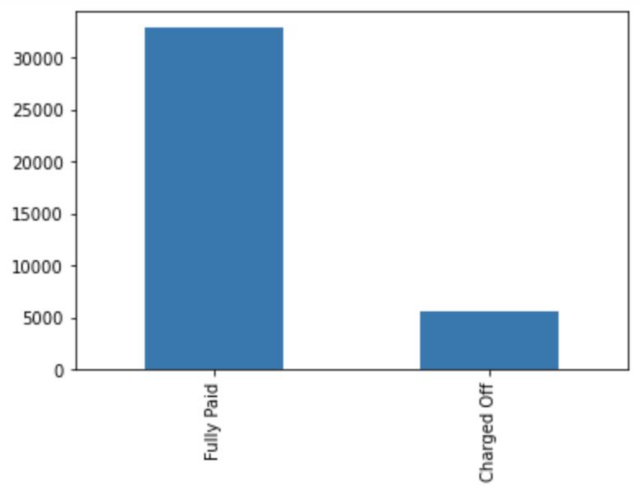
Exploring target feature (loan_status) distribution

```
[68]: df.loan_status.value_counts(1)
      Last executed at 2022-05-11 20:35:54 in 21ms

[68]: Fully Paid     0.854349
      Charged Off    0.145651
      Name: loan_status, dtype: float64

[69]: df.loan_status.value_counts().plot(kind = 'bar') #
      Last executed at 2022-05-11 20:36:13 in 148ms

[69]: <AxesSubplot:>
```



Data Analysis
(Segmented Univariate)

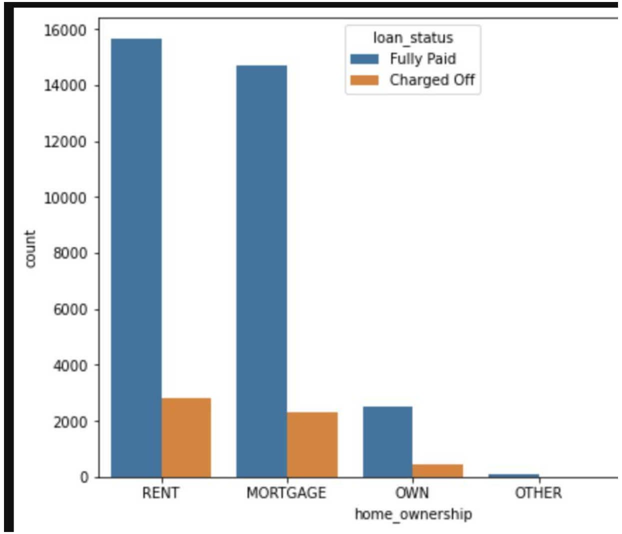
With segmented univariate analysis we were able to understand how different features impact our target variables(loan_status)

Data Analysis
(Segmented Univariate) - Home Ownership & Loan Status

```
sc_ho = segmented_comparison('home_ownership'
Last executed at 2022-05-09 17:53:11 in 47ms
```

	home_ownership	loan_status	count	ratio
0	MORTGAGE	Charged Off	2327	0.136713
1	MORTGAGE	Fully Paid	14694	0.863287
2	NONE	Fully Paid	3	1.000000
3	OTHER	Charged Off	18	0.183673
4	OTHER	Fully Paid	80	0.816327
5	OWN	Charged Off	443	0.148908
6	OWN	Fully Paid	2532	0.851092
7	RENT	Charged Off	2839	0.153626
8	RENT	Fully Paid	15641	0.846374

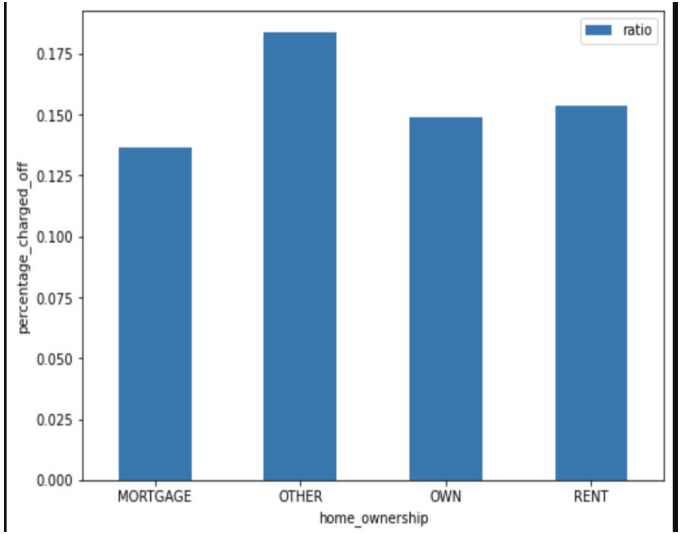
Home Ownership & Loan status distribution



Home Ownership & Loan status distribution

Data Analysis
(Segmented Univariate) - Home Ownership & Loan Status

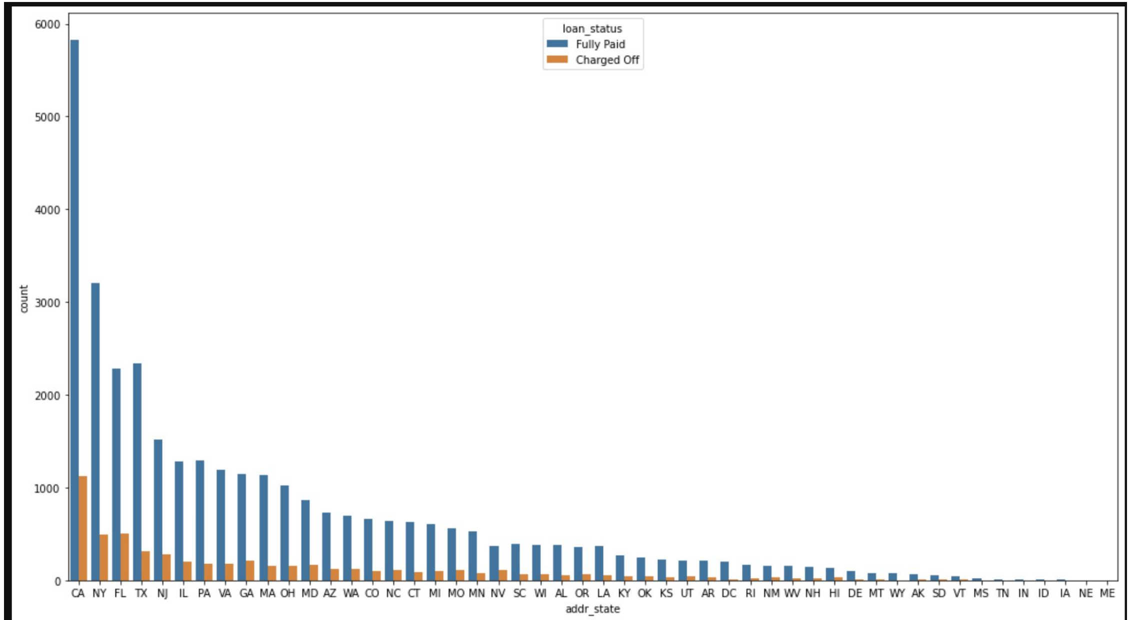
Default percentage by Home Ownership



<u>we can clearly see from this analysis that users with `other` home ownership default much more often, at ~18% vs. a median of ~14% in other (own, rent, mortgage) ownership types</u>

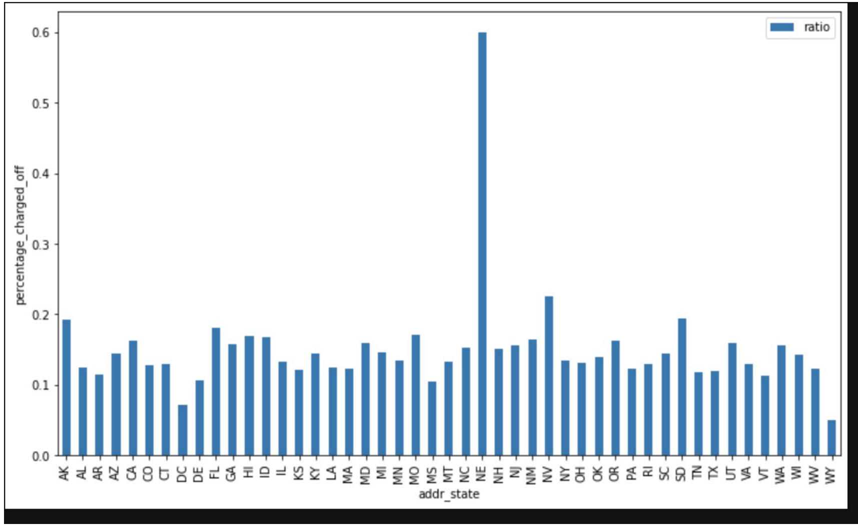
Data Analysis
(Segmented Univariate) - Address State & Loan Status

Address State & Loan status distribution



Data Analysis (Segmented Univariate) - Address State & Loan Status

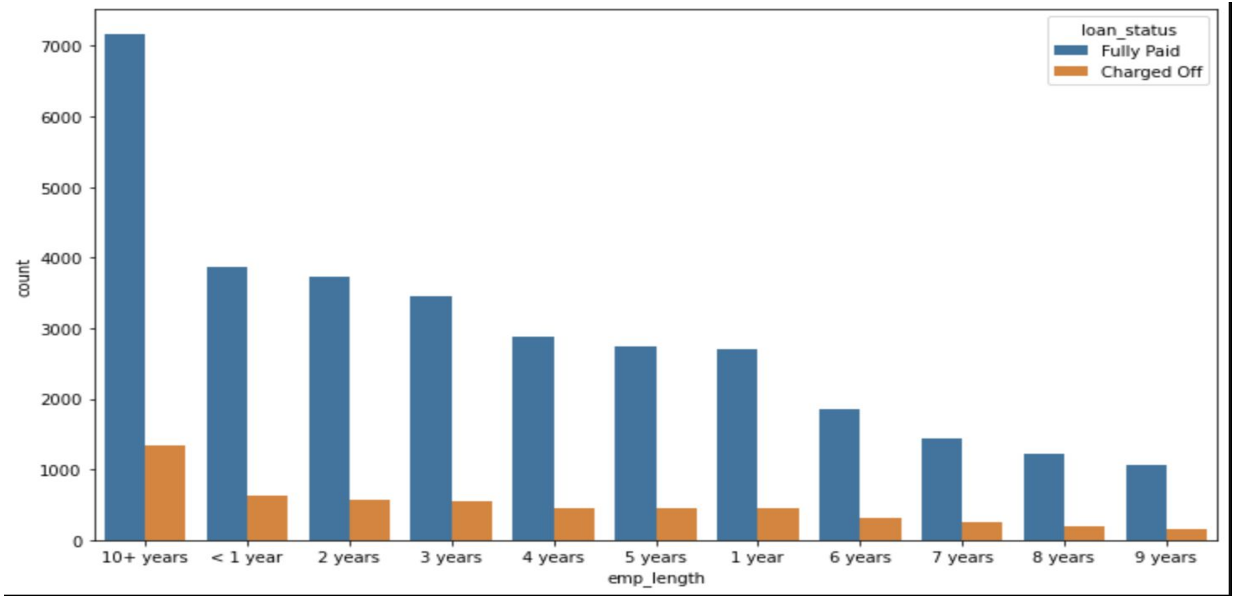
Default percentage by Address State



<u>we can see here that users from state `NE` have a whopping 60% of defaulting. It is worth mentioning only a total of 5 loans were issued in this state, so we don't have big enough sample size. The next worse performing state is `NV` with a ~22% default rate</u>

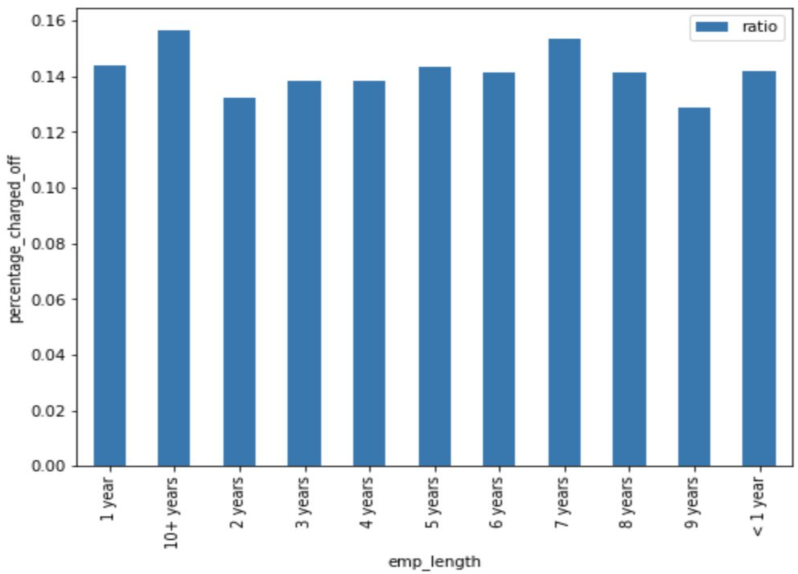
Data Analysis (Segmented Univariate) - Emp. Length & Loan Status

Emp. Length
&
Loan status
distribution



Data Analysis
(Segmented Univariate) - Emp. Length & Loan Status

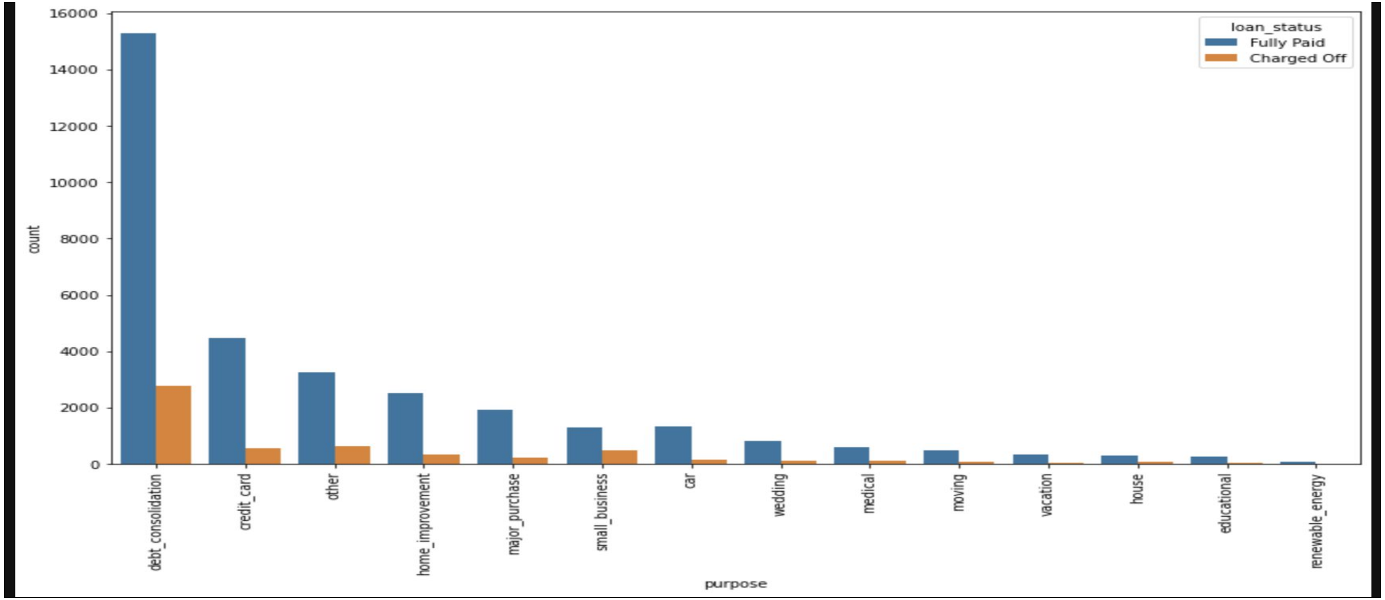
Emp. Length
&
Default
percentage



There is not much of a variance when it comes to a users work experience and they defaulting to conclude anything. Although it seems users with 10+ years of work exp. default the most

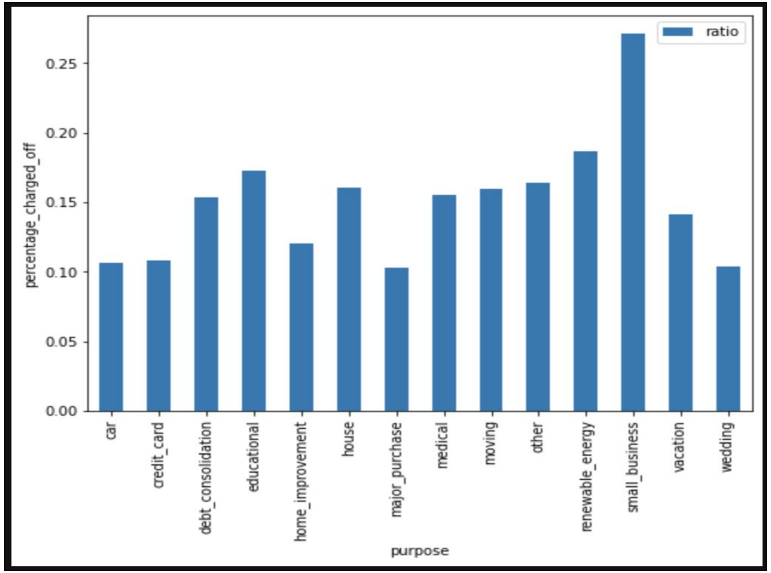
Data Analysis (Segmented Univariate) - Purpose & Loan Status

Purpose & Loan status distribution



Data Analysis
(Segmented Univariate) - Purpose & Loan Status

Default
percentage
by Purpose

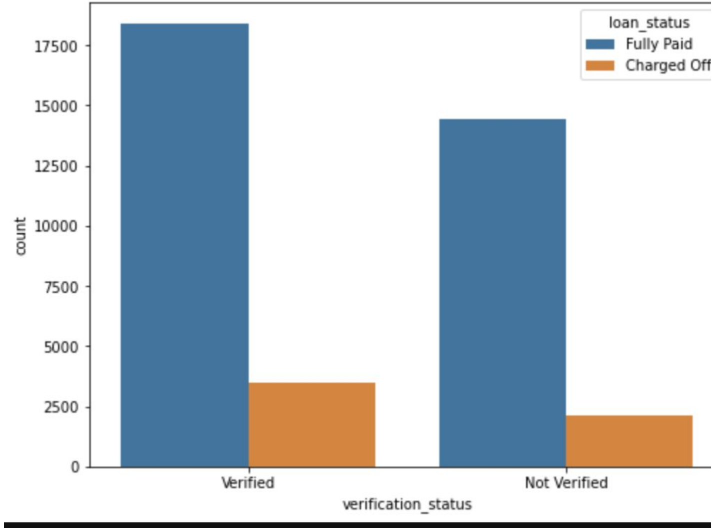


<u>we can see that loans for `small_business` are the least secure and more likely to default than others</u>

Data Analysis (Segmented Univariate) - verification status & Loan Status

	verification_status	loan_status	count	ratio
0	Not Verified	Charged Off	2122	0.127955
1	Not Verified	Fully Paid	14462	0.872045
2	Verified	Charged Off	3478	0.159074
3	Verified	Fully Paid	18386	0.840926

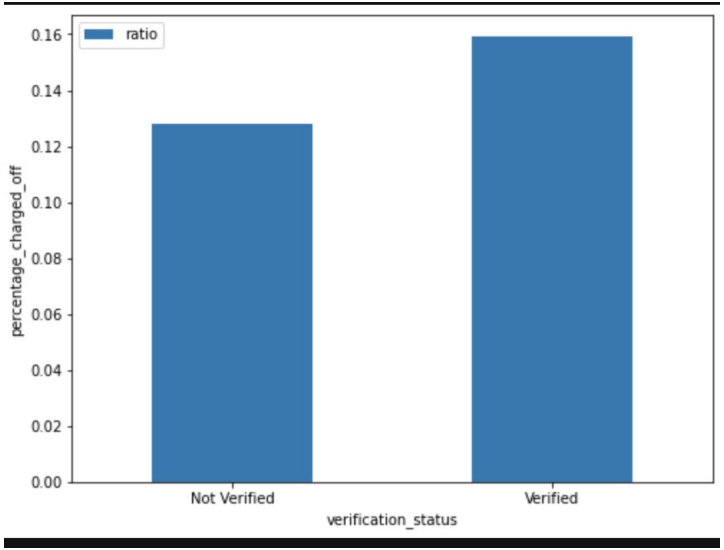
verification_status & Loan status distribution



Verification status & Loan status distribution

Data Analysis (Segmented Univariate) - Verification Status & Loan Status

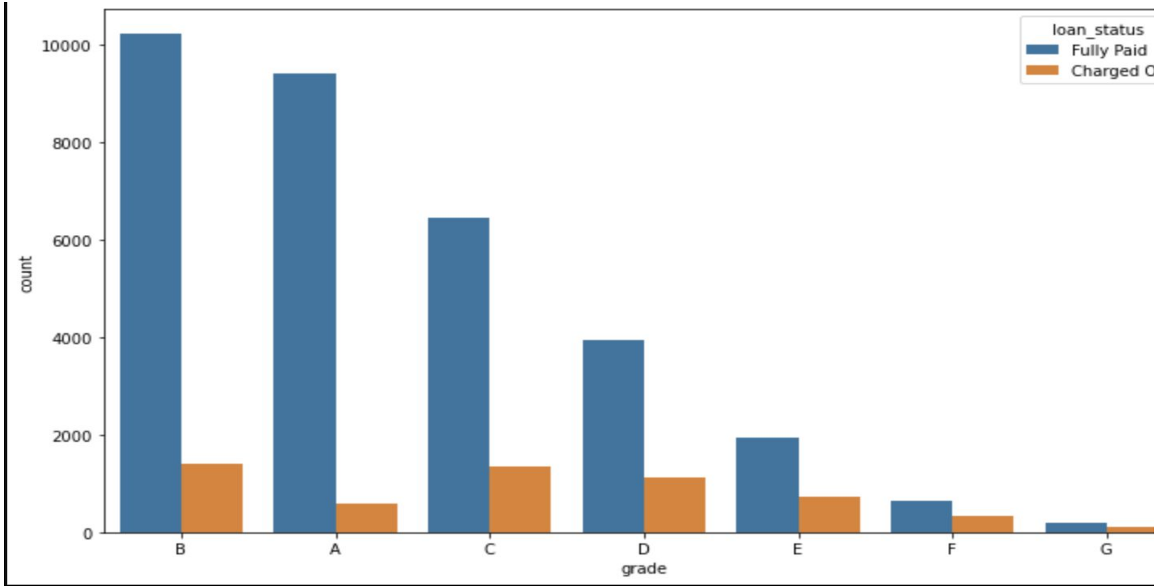
Default percentage by verification Status



we can see here that surprisingly users with verified income source are more likely to default than non-verified users

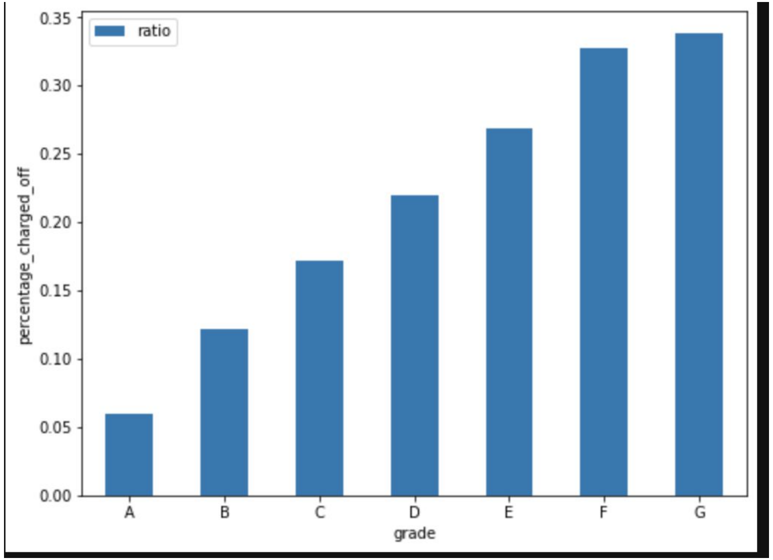
Data Analysis (Segmented Univariate) - Grade & Loan Status

Grade & Loan status distribution



Data Analysis
(Segmented Univariate) - Grade & Loan Status

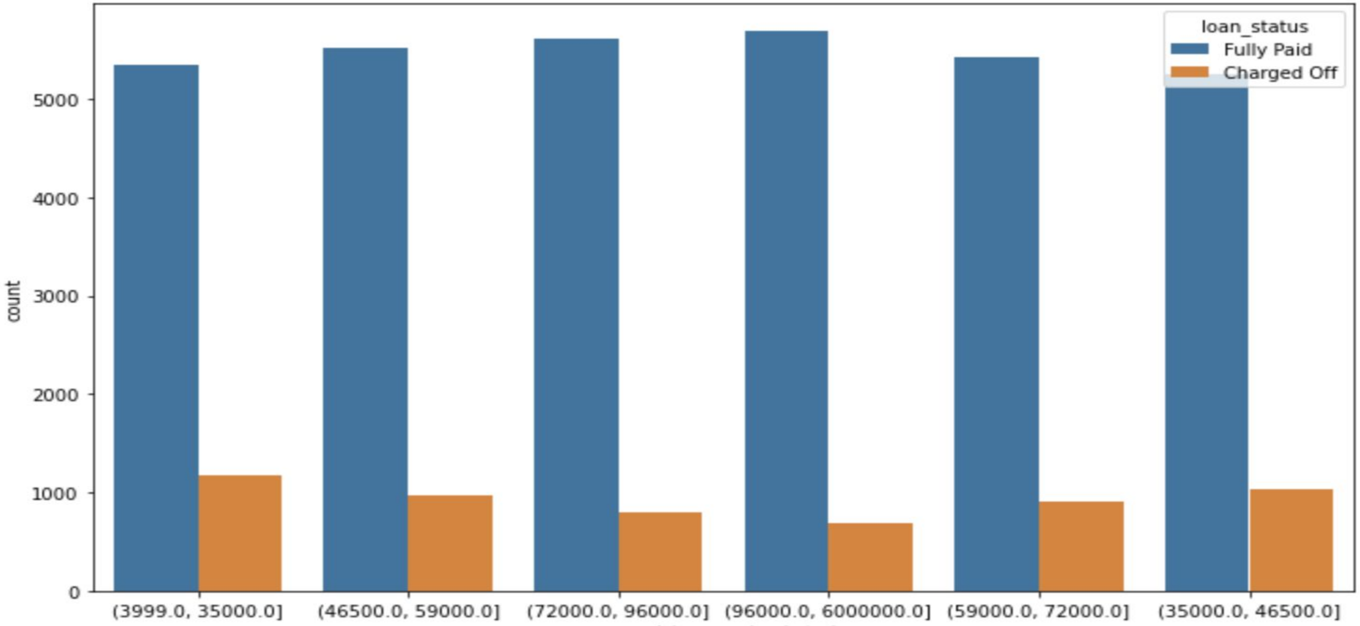
Default percentage by Grade



As expected grade plays a major role in whether a user will default or not. The default percentage seems to keep on increasing as the grade increases

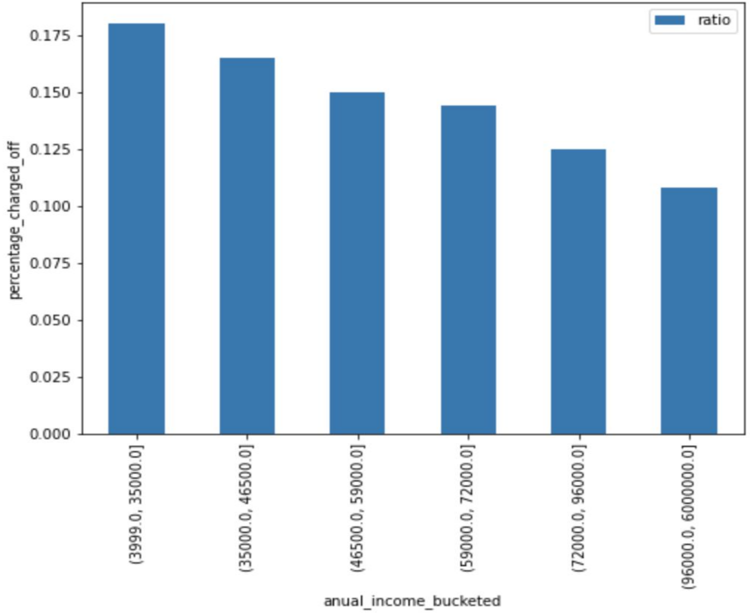
Data Analysis (Segmented Univariate) - Annual income & Loan Status

Annual income & Loan status distribution



Data Analysis
(Segmented Univariate) - Annual Income & Loan Status

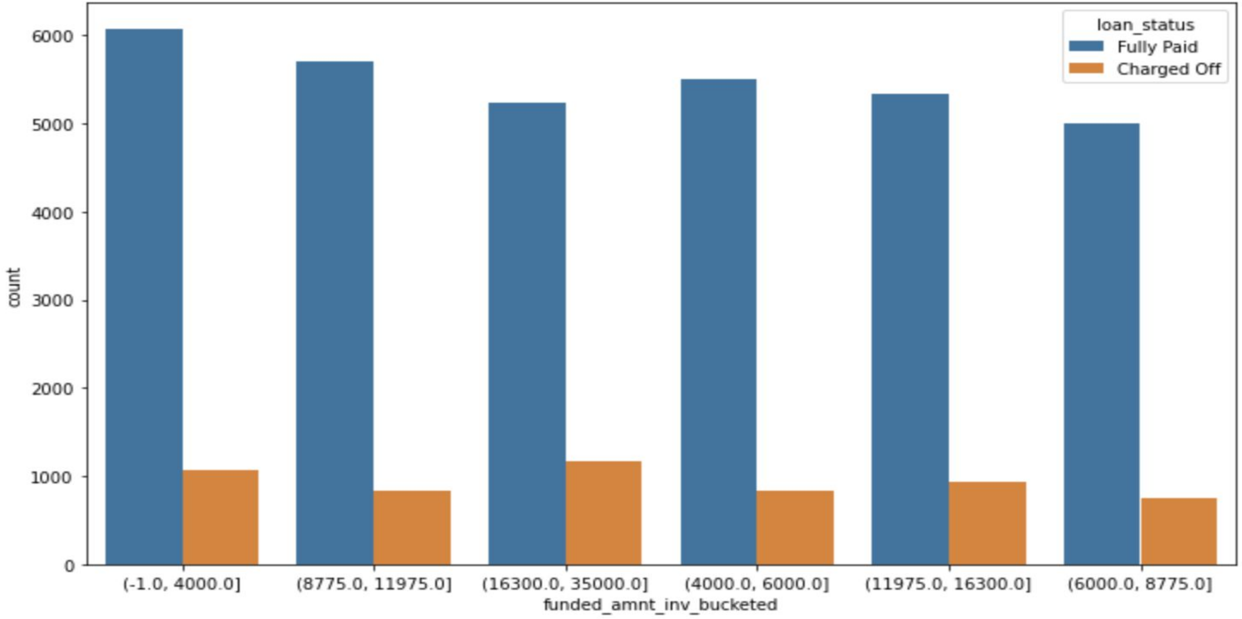
Default percentage by Annual Income



As seen above Lower income users are more likely to default than higher income users

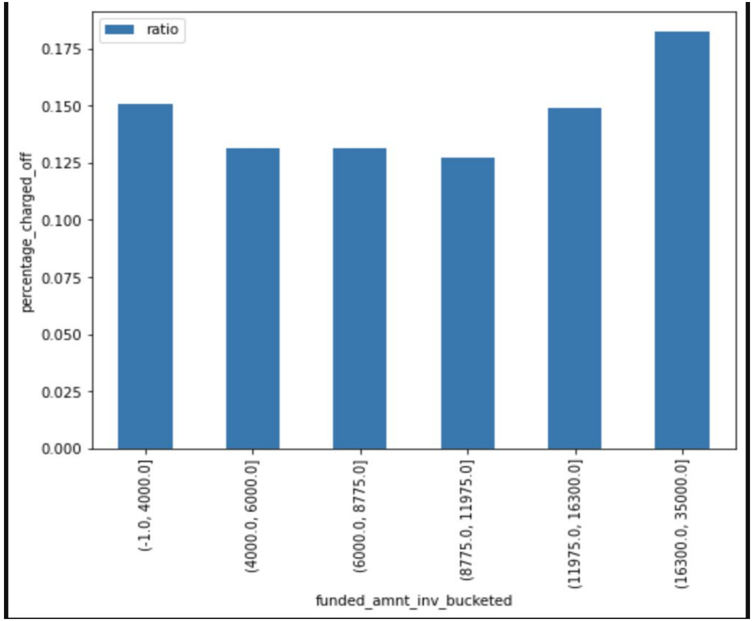
Data Analysis
(Segmented Univariate) - Loan amount & Loan Status

Loan amount
&
Loan status
distribution



Data Analysis
(Segmented Univariate) - Loan Amount & Loan Status

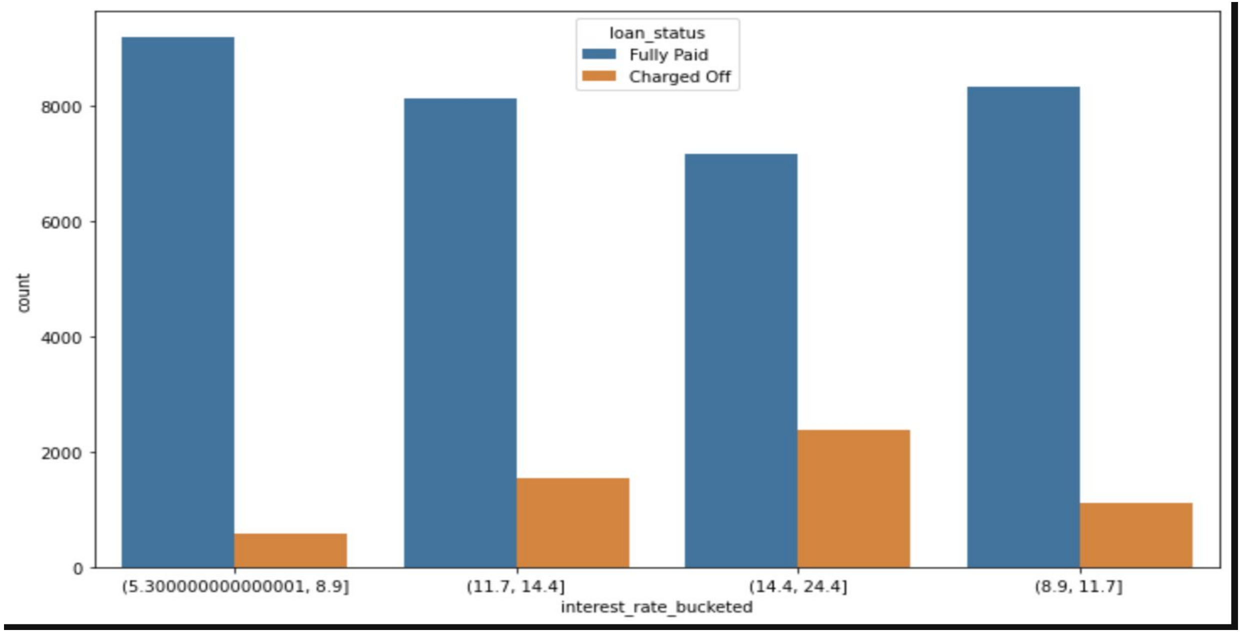
Default percentage by loan amount



As the loan amount increases, the cahnces of default also increases

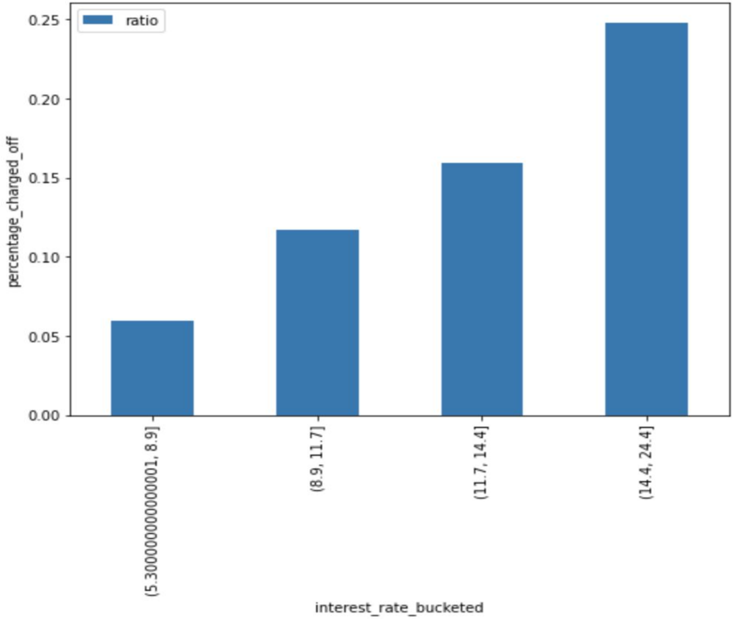
Data Analysis
(Segmented Univariate) - Interest rate & Loan Status

Interest Rate
&
Loan status
distribution



Data Analysis (Segmented Univariate) - Interest rate & Loan Status

Default percentage by interest rate

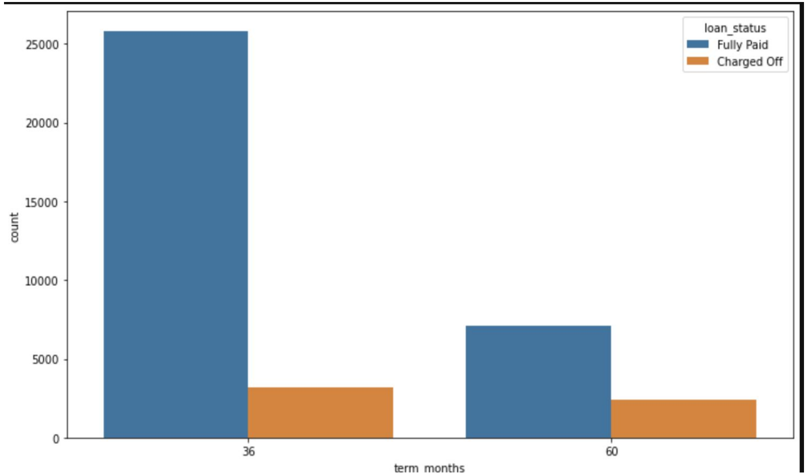


From the above information we can conclude that users who are charged a higher interest rate are way more lilely to default than users being charged a lower rate

Data Analysis (Segmented Univariate) - Term & Loan Status

	term_months	loan_status	count	ratio
0	36	Charged Off	3200	0.110471
1	36	Fully Paid	25767	0.889529
2	60	Charged Off	2400	0.253138
3	60	Fully Paid	7081	0.746862

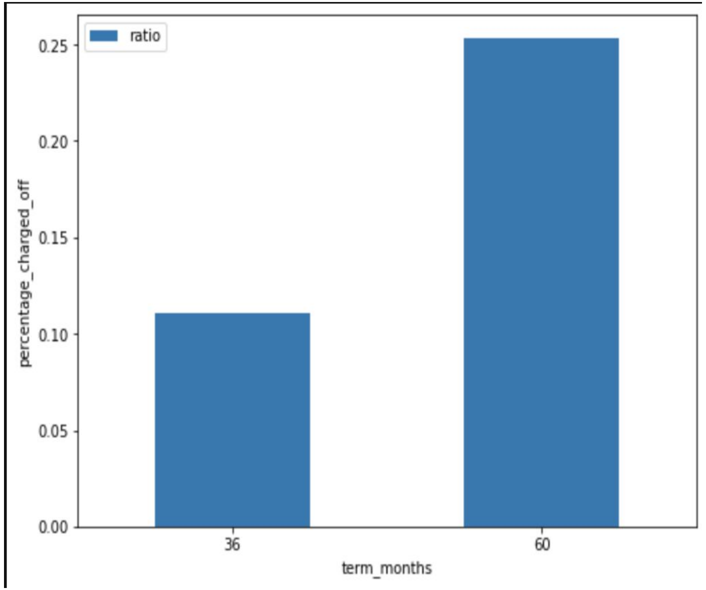
Term & Loan status distribution



Term & Loan status distribution

Data Analysis
(Segmented Univariate) - Term & Loan Status

Default percentage by term

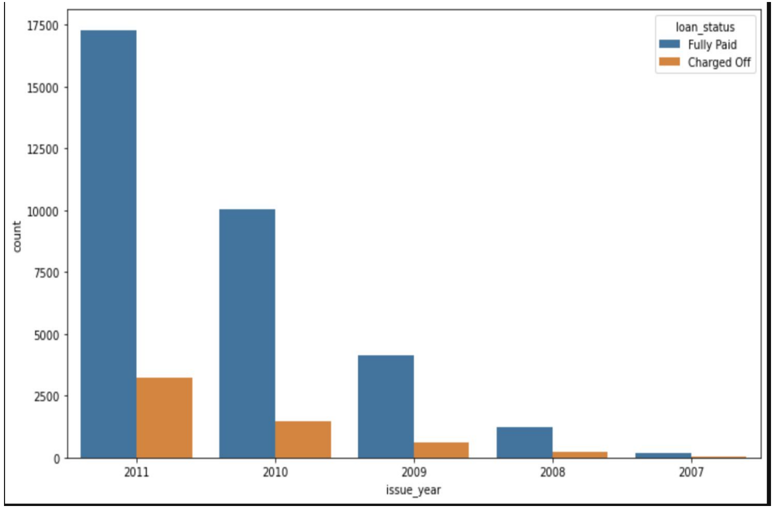


<u>From the above information we can conclude that users with 60 months term are more highly likely to default</u>

Data Analysis
(Segmented Univariate) - Issue Year & Loan Status

	issue_year	loan_status	count	ratio
0	2007	Charged Off	45	0.180000
1	2007	Fully Paid	205	0.820000
2	2008	Charged Off	220	0.153417
3	2008	Fully Paid	1214	0.846583
4	2009	Charged Off	594	0.125954
5	2009	Fully Paid	4122	0.874046
6	2010	Charged Off	1485	0.128772
7	2010	Fully Paid	10047	0.871228
8	2011	Charged Off	3256	0.158705
9	2011	Fully Paid	17260	0.841295

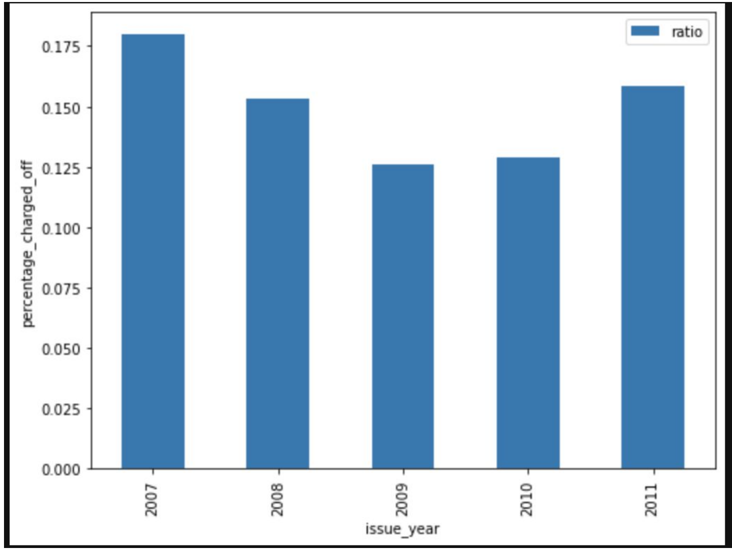
Issue year & Loan status distribution



Issue year & Loan status distribution

Data Analysis
(Segmented Univariate) - Issue Year & Loan Status

Default percentage by Issue Year



<u>we can clearly see above that most defaults were in the year 2007, this is probably due to the recession</u>

Data Analysis (Multiivariate)

from the above correlation matrix we can conclude that loan amount and installments have a hhigh correlation, which is quite obvious. no other features are highly co-related

Conclusion & Recommendations

Conclusion:

From the Analysis we performed we can say that following are the driving features towards loan default:

- **Grade** (user credit grade)
- **int_rate** (interest rate %)
- **term** (loan duration)
- **home_ownership** (user home ownership status)
- **Purpose** (loan purpose)
- **funded_amnt_inv** (loan amount)
- **annual_income** (users annual income)

Conclusion & Recommendations

<u>Recommendations:</u>

Since Now we know what are the driving features that leads to a loan defaulting, we can recommend lending club the following

- **Users with higher credit grades(A, B, C) are less likely to default and users with low grades(E, F, G) are very likely to default**
- **If the interest rate is kept under 10%, there's a good chance user won't default**
- **It is recommended that loan term be kept to 36 months only**
- **It'd be better if loans are not given to individuals with home ownership status as other/unknown.**
- **Loans for funding small businesses are very risky. Loans taken for purchasing big equipments like cars seems to be less risky.**
- **As the loan amount increases, the chances of default also increases**
- **Users with low annual income have a high chance to default, so appropriate actions needs to be taken while providing loans to low income users**

Thank You

Case Study By - Rishabh Vij