ANALYTIXLABS

Data Engineering: Final Projects

Data Engineering Projects

Project1: Retail Sales Analytics

Project1: Retail Sales Analytics

Overview:

The objective of the project to illustrate retail analytics using an online retail dataset containing transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. This dataset is used to demonstrate an end-to-end retail analytic use case on the Hadoop Data Platform distribution:

- Data ingestion and cleansing using Apache Pig/Hive
- SQL on Hadoop using Hive
- Analytics and visualization using Hive/SparkSQL/Tableau

Project1: Retail Sales Analytics

Data set:

The original Online Retail data set is available to download on the [UCI Machine Learning Repository] (https://archive.ics.uci.edu/ml/datasets/Online+Retail). It has been converted using Microsoft Excel to a tab delimited file available for convenience. The fields in the data as follows.

✓InvoiceNo - integer - Transaction Number
✓StockCode - character - SKU Code (Product Code)
✓Description - character - Product Description
✓Quantity - int - Quantity ordered
✓InvoiceDate - character - Transaction Data
✓UnitPrice - float- Price per unit quantity
✓CustomerID - character - Customer ID
✓Country - character - Customer location

Data Download Link

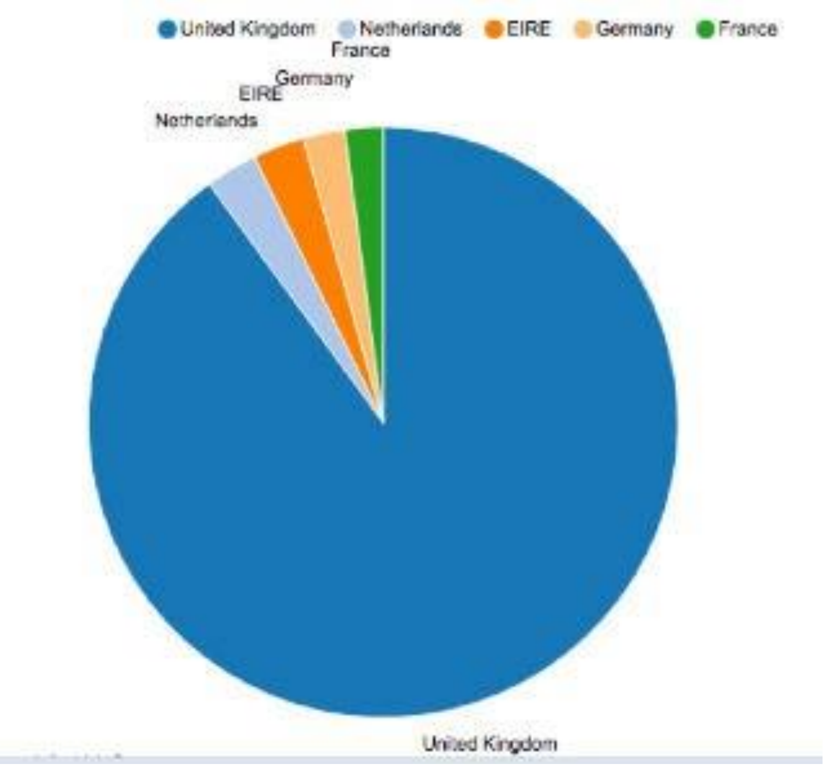
Project1: Retail Sales Analytics

Analysis:

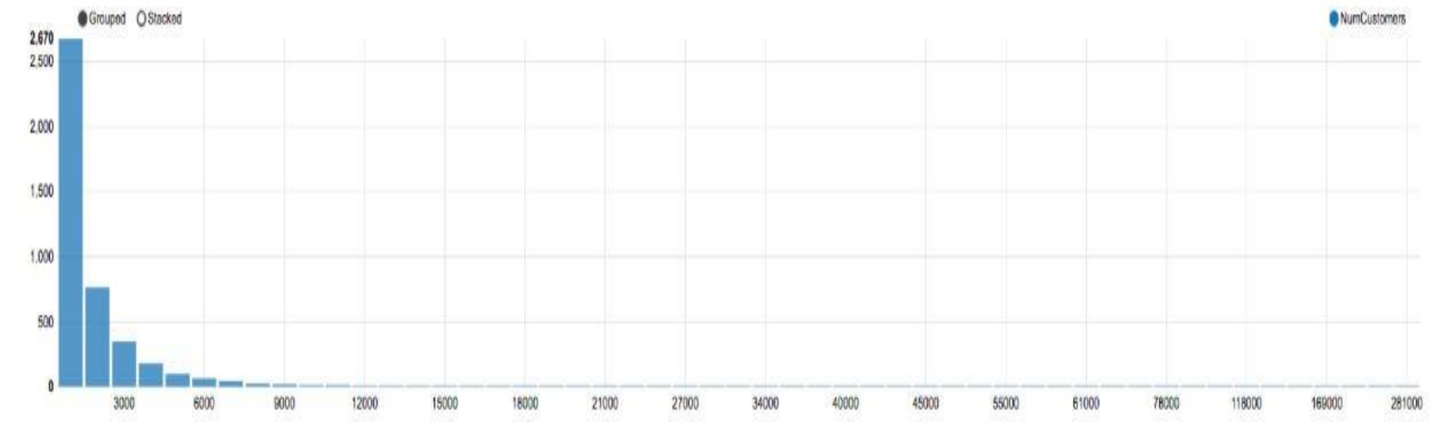
1. Revenue Aggregate By Country for top 5 countries
2. Sales Metrics like NumCustomers, NumTransactions, AvgNumItems, MinAmtperCustomer, MaxAmtperCustomer, AvgAmtperCustomer, StdDevAmtperCustomer etc. .. by country for top 5 countries
3. Daily Sales Activity like NumVisits, TotalAmt etc… per POSIX day of the year
4. Hourly sales Activity like NumVisits, TotalAmt etc… per hour of day
5. Basket size distribution (Note: Basket size = number of items in a transaction) (in this questions, we would like to know that, number of transactions by each basket size i.e number of transactions with 3 size, number of transactions with 4 size etc…
6. Top 20 Items sold by frequency
7. Customer Lifetime Value distribution by intervals of 1000’s (Customer Life time Value = total spend by customer in his/her tenure with the company) (In this question, we would like to calculate how many customers with CLV between 1-1000, 1000-2000 etc.). Please note that we don’t want calculate bins manually and it required to create bins dynamically.

Sample Outputs

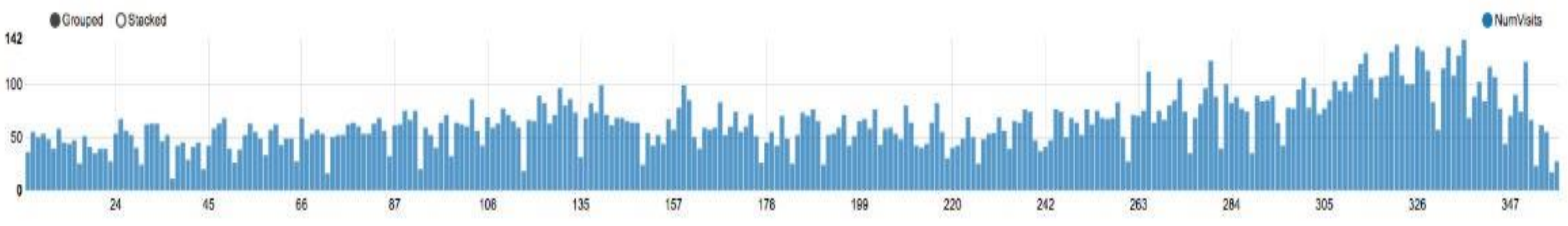
Revenue by country



Distribution of Customer Life time value intervals of 1000’s for the customer base

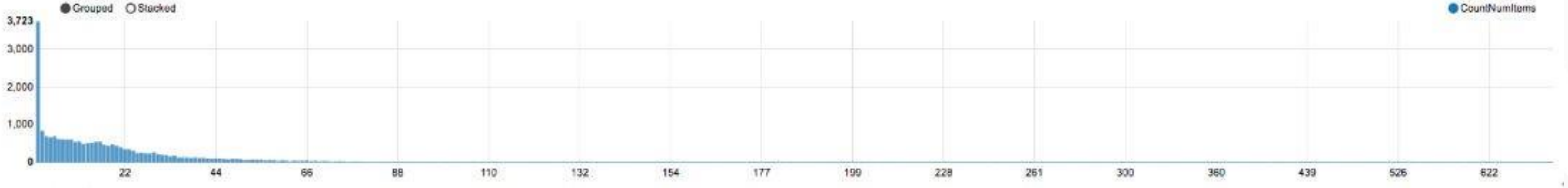


Daily Sales Activity

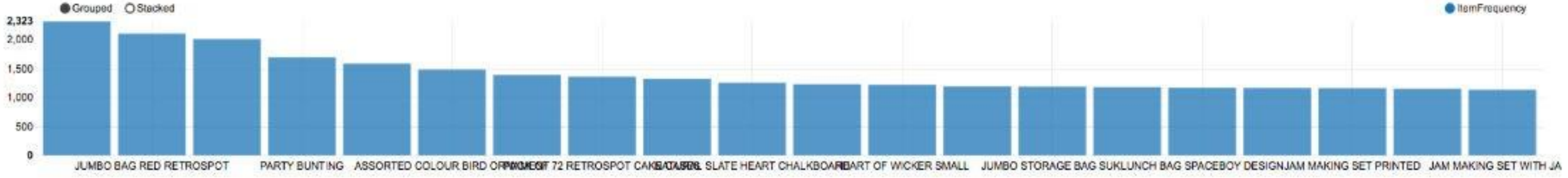


Sample Outputs

Basket Size Distribution



Top 20 Items sold by frequency



Project2: Bank Loan Analysis

Project2: Lending club Loan Analysis

About Company:

Lending Club (LC) is one of the largest online credit marketplace, facilitating personal loans, business loans, and financing for elective medical procedures. Borrowers access lower interest rate loans through a fast and easy online or mobile interface. Investors provide the capital to enable many of the loans in exchange for earning interest. LC operate fully online with no branch infrastructure, and use technology to lower cost and deliver an amazing experience. LC pass the cost savings to borrowers in the form of lower rates and investors in the form of attractive returns. LC is transforming the banking system into a frictionless, transparent and highly efficient online marketplace, helping people achieve their financial goals everyday.

How LC Operates:

✓ As a personal loan or business loan borrower, you can get an instant quote in minutes with no impact to your credit score. Once you select an offer, you can watch as funds are committed by investors who are choosing to invest in you and your success.

✓ If you're investing, you can open an account in minutes and build a portfolio of hundreds or thousands of loans made to quality borrowers. You'll receive monthly payments of principal and interest, which you can withdraw or reinvest.

✓ If you're looking for medical financing, you can apply online or through our network of more than 10,000 providers across the country.

✓ No matter what kind of loan you're interested in, everything is done online, so the whole process is fast, convenient and private.

✓ All loans facilitated by Lending Club are issued by a bank and subject to the same consumer protection, fair lending, and disclosure requirements as any other bank loan.

Project2: Lending Club Loan Analysis

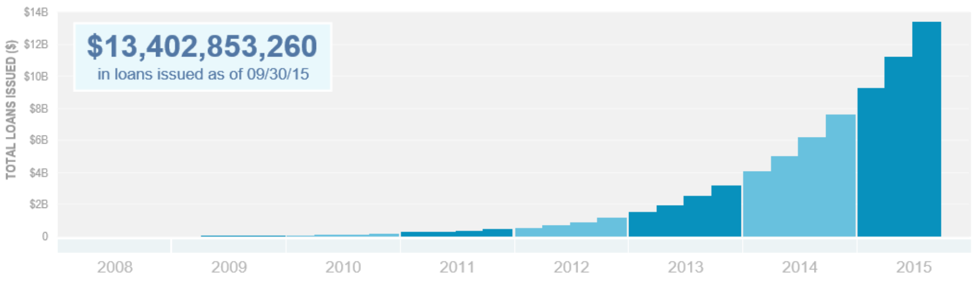
Business Objective:

The objective of this analysis to understand loans performance in different dimensions. In order to understand loans performance the company would like to do following analysis using available data.

- ✓ **Total Loan issuance by yearly & quarterly and calculate growth rate by quarter on quarter and year on year**
- ✓ **Percentage of loans based on reported loan purpose. (Note:** Loan purpose describes the reported intent of borrowers from the most recent completed quarter and may not reflect actual usage. Investors should rely on loan grades rather than loan purpose)
- ✓ **Loan Issuance by state – classify the states based on loan issuance by $50+ MM, $25-50 MM, $10-25 MM and $0-10 MM**
- ✓ **Find the last quarter average interest rates by different term loans and overall**
- ✓ **Find the historical returns by loan grade(Historical performance by grade for all issued loans) and overall**
- ✓ **Find the historical average interest rates by loan terms and loan grades (also for overall)**
- ✓ **What is percentage of loans by different loan grades by each year and loan term level (also for overall)**
- ✓ **What is the loan performance details by different loan grades and overall**
- ✓ **Find Net Annualized returns by vintage by different loan grades and different loan terms (also for overall**
- ✓ **What is loan status migration over 9 months (Net Charge offs: 120+days delinquency)**

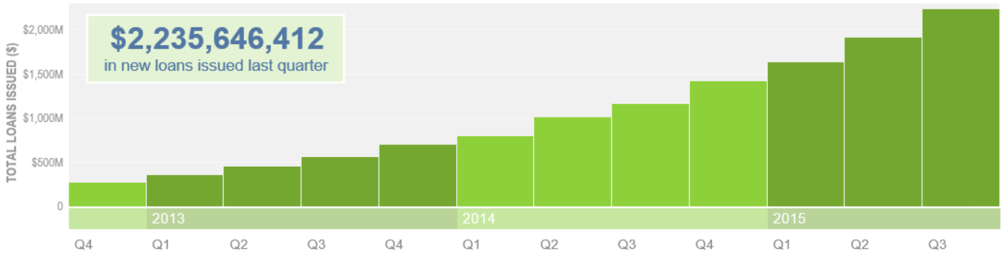
Project2: Lending Club Loan Analysis -Analysis samples

TOTAL LOAN ISSUANCE



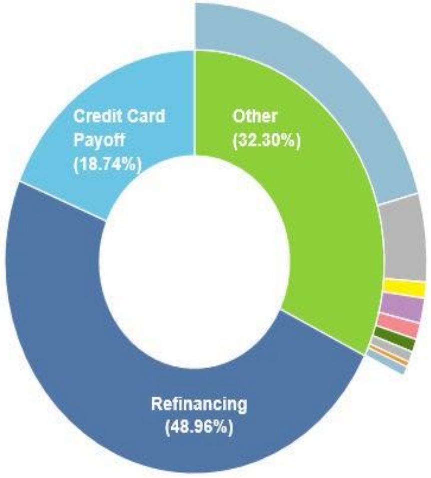
$13,402,853,260
in loans issued as of 09/30/15

QUARTERLY LOAN ISSUANCE



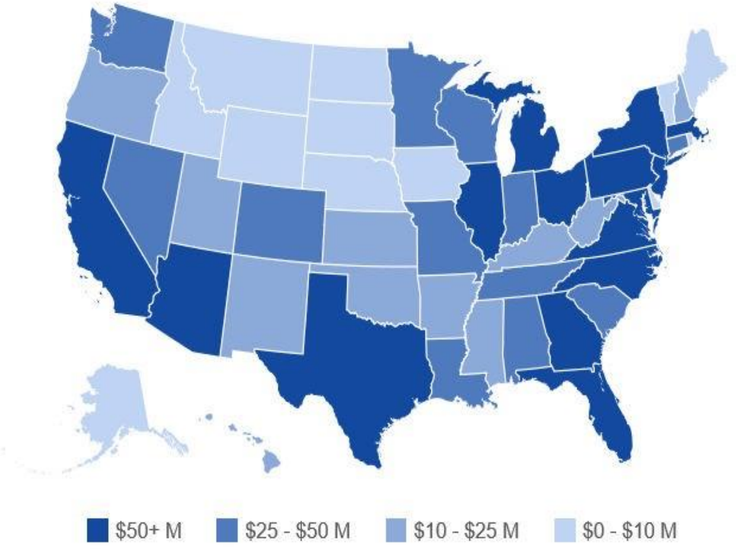
$2,235,646,412
in new loans issued last quarter

REPORTED LOAN PURPOSE

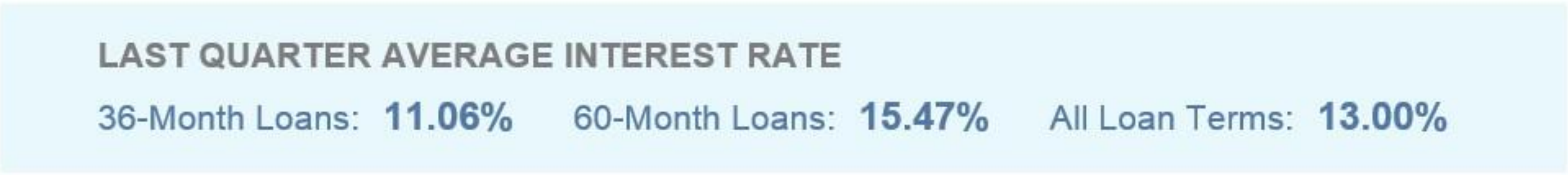


67.70% of Lending Club borrowers report using their loans to refinance existing loans or pay off their credit cards as of 09/30/15.[1]

LOAN ISSUANCE BY STATE



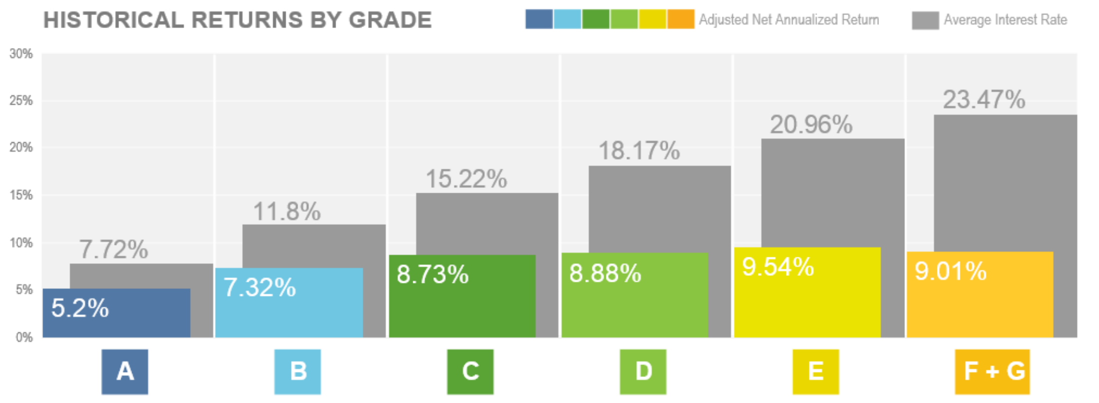
Most recent reported quarter ending 09/30/15

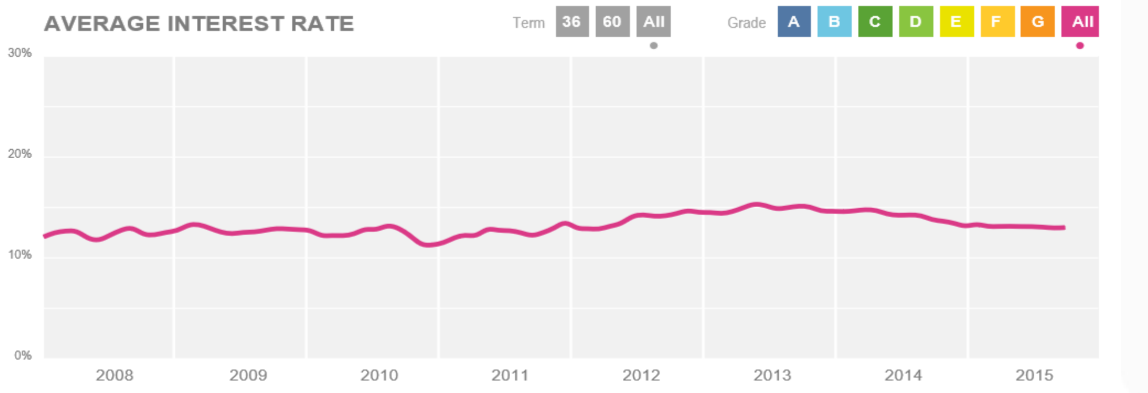


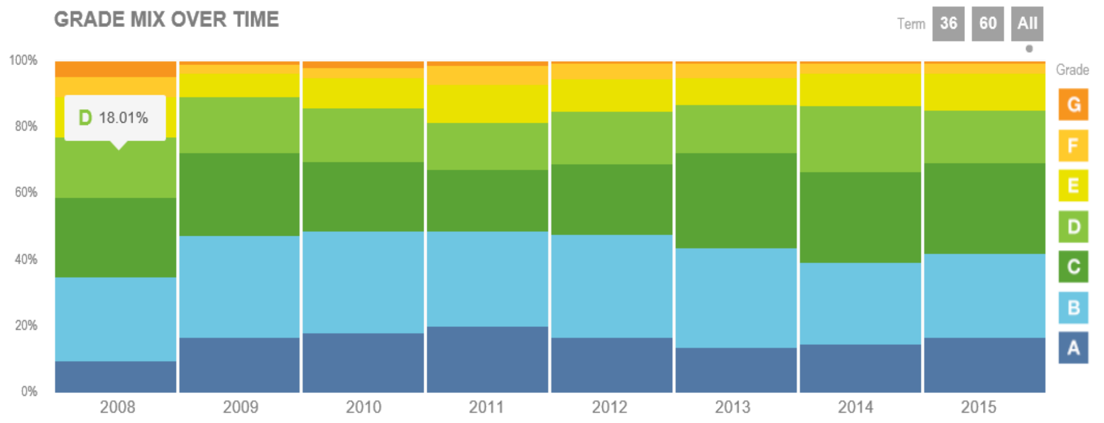
LAST QUARTER AVERAGE INTEREST RATE

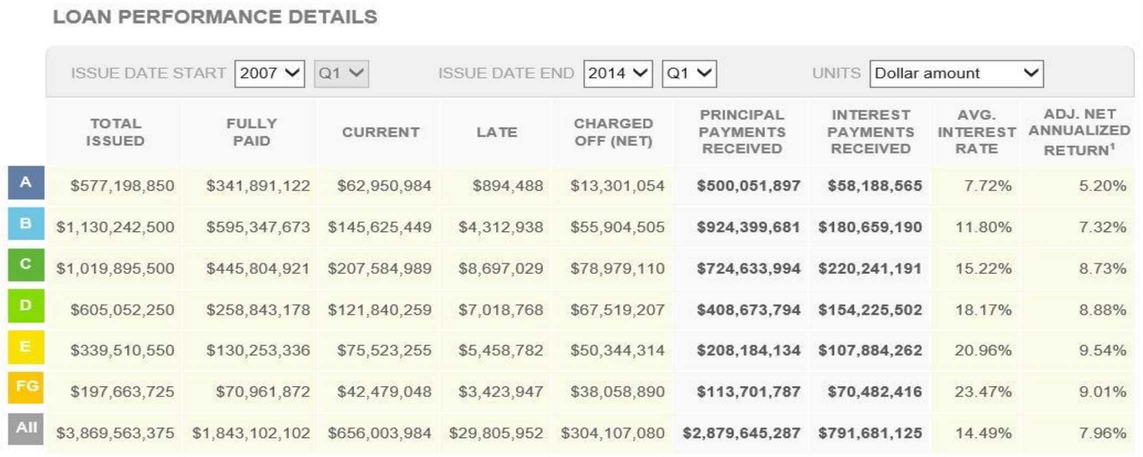
36-Month Loans: **11.06%** 60-Month Loans: **15.47%** All Loan Terms: **13.00%**

Project2: Lending Club Loan Analysis -Analysis samples







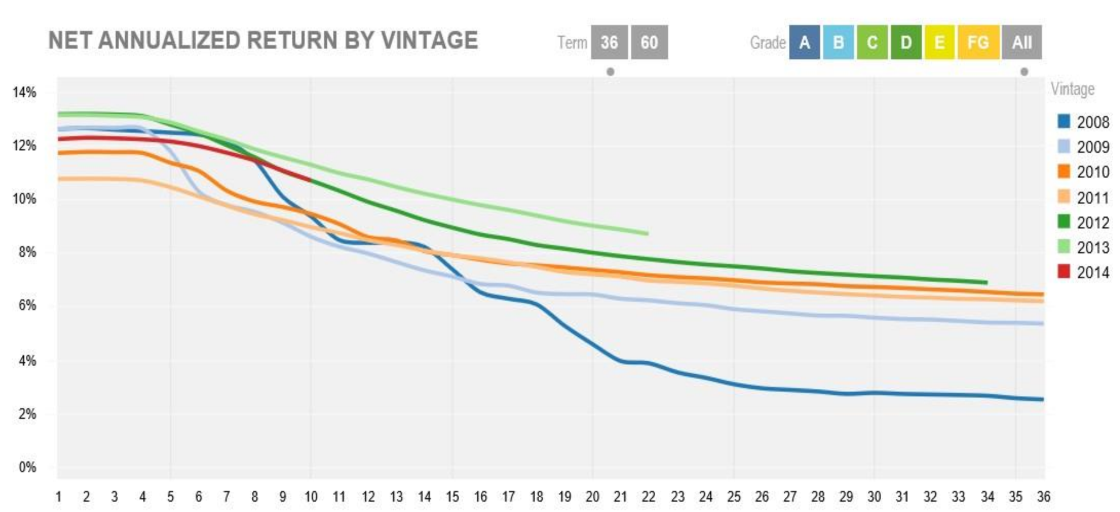


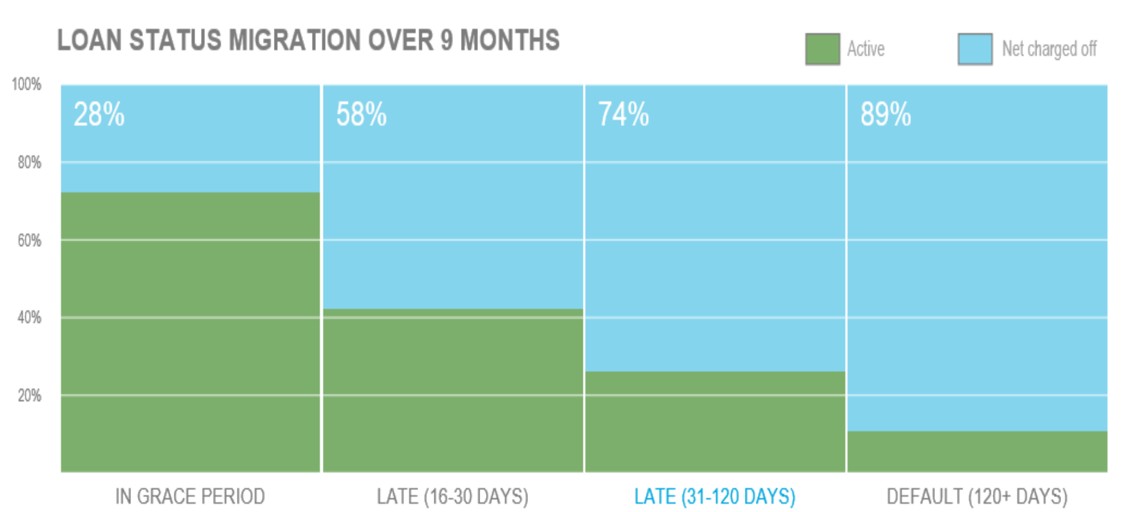
LOAN PERFORMANCE DETAILS

ISSUE DATE START 2007 Q1 ISSUE DATE END 2014 Q1 UNITS Dollar amount

	TOTAL ISSUED	FULLY PAID	CURRENT	LATE	CHARGED OFF (NET)	PRINCIPAL PAYMENTS RECEIVED	INTEREST PAYMENTS RECEIVED	AVG. INTEREST RATE	ADJ. NET ANNUALIZED RETURN[1]
A	$577,198,850	$341,891,122	$62,950,984	$894,488	$13,301,054	**$500,051,897**	**$58,188,565**	7.72%	5.20%
B	$1,130,242,500	$595,347,673	$145,625,449	$4,312,938	$55,904,505	**$924,399,681**	**$180,659,190**	11.80%	7.32%
C	$1,019,895,500	$445,804,921	$207,584,989	$8,697,029	$78,979,110	**$724,633,994**	**$220,241,191**	15.22%	8.73%
D	$605,052,250	$258,843,178	$121,840,259	$7,018,768	$67,519,207	**$408,673,794**	**$154,225,502**	18.17%	8.88%
E	$339,510,550	$130,253,336	$75,523,255	$5,458,782	$50,344,314	**$208,184,134**	**$107,884,262**	20.96%	9.54%
FG	$197,663,725	$70,961,872	$42,479,048	$3,423,947	$38,058,890	**$113,701,787**	**$70,482,416**	23.47%	9.01%
All	$3,869,563,375	$1,843,102,102	$656,003,984	$29,805,952	$304,107,080	**$2,879,645,287**	**$791,681,125**	14.49%	7.96%

Project2: Lending Club Loan Analysis -Analysis samples





Note: The Analysis samples are just for your reference. The numbers may not match with your numbers.
You can integrate tableau/jasper reports/excel2013 to come up with similar reports. You can also provide excel based results once you complete the analysis in Hadoop.

Project2: Lending Club Loan Analysis

Input Data

Loans Data(4 Files):
These files contain complete loan data for all loans issued through the time period stated, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter.

Declined Loan Data(3 Files):
These files contain the list and details of all loan applications that did not meet Lending Club's credit underwriting policy.

Detailed Data Dictionary:
The Data Dictionary includes definitions for all the data attributes included in the Historical data file and the In Funding data file.

Data Download Link

Project2: Lending Club Loan Analysis

Expected outcome:

As per basic expectations, you need to come up with the analysis as per the business objective. Please refer analysis samples to get understand the questions

The data is very rich. You can come with many analysis related to

- ✓ Loans performance
- ✓ Understanding on rejection of loans
- ✓ Investor performance
- ✓ Delinquency rates by each quarter or year
- ✓ Prepayment rates by each quarter or year
- ✓ Cumulative charge offs

Project3: Analyzing Machine & Sensor Data

Project3: Analyzing machine & sensor data

Sensor Data:

A sensor is a device that measures a physical quantity and transforms it into a digital signal. Sensors are always on, capturing data at a low cost, and powering the “Internet of Things.”

Potential Uses of Sensor Data:

Sensors can be used to collect data from many sources, such as:

- ✓ To monitor machines or infrastructure such as ventilation equipment, bridges, energy meters, or airplane engines. This data can be used for predictive analytics, to repair or replace these items before they break.
- ✓ To monitor natural phenomena such as meteorological patterns, underground pressure during oil extraction, or patient vital statistics during recovery from a medical procedure.

This case study is about how to refine data from heating, ventilation, and air conditioning (HVAC) systems using the Cloudera Data Platform, and how to analyze the refined sensor data to maintain optimal building temperatures.

Project3: Analyzing machine & sensor data

Input Data

In this case study, we will focus on sensor data from building operations. Specifically, we will refine and analyze the data from Heating, Ventilation, Air Conditioning (HVAC) systems in 20 large buildings around the world

In order to perform analysis, we will use the below data as input data.

✓ **HVAC.csv** – contains the targeted building temperatures, along with the actual (measured) building temperatures. The building temperature data was obtained using Apache Flume. Flume can be used as a log aggregator, collecting log data from many diverse sources and moving it to a centralized data store. In this case, Flume was used to capture the sensor log data, which we can now load into the Hadoop Distributed File System (HFDS).

✓ **building.csv** – contains the "building" database table. Apache Sqoop can be used to transfer this type of data from a structured database into HFDS.

Data Download Link

Project3: Analyzing machine & sensor data

High Level Steps:

1. Download and extract the sensor data files(HVAC.csv & building.csv).
2. Load the sensor data into the HDFS.
3. Create HVAC & Building tables in Hive or Pig
4. Using Hive/pig/impala to refine the sensor data.

5. Calculate three variables (temp_diff, temprange, extremetemp) in Hvac table

   - ✓ Temp_didff = actual temperature – target temperature
   - ✓ temprange column indicates whether the actual temperature was:
     - ✓ NORMAL – within 5 degrees of the target temperature.
     - ✓ COLD – more than five degrees colder than the target temperature.
     - ✓ HOT – more than 5 degrees warmer than the target temperature.
   - ✓ Extrememetemp: If the temperature is outside of the normal range, **extremetemp** is assigned a value of 1; otherwise its value is 0.

6. Create final file by joining the two tables and perform the analysis

Project3: Analyzing machine & sensor data

Expected Output:

We would like to accomplish three goals with this data:

- ✓ Reduce heating and cooling expenses.
- ✓ Keep indoor temperatures in a comfortable range between 65-70 degrees.
- ✓ Identify which HVAC products are reliable, and replace unreliable equipment with those models.

These analysis will be very helpful for facilities department to initiate data-driven strategies to reduce energy expenditures and improve employee comfort.

Analysis need to be performed:

1. Data visualization/analysis by mapping the buildings that are most frequently outside of the optimal temperature range. Calculate count of extremetemp (i.e where the temperature was more than five degrees higher or lower than the target temperature) by each country and temprange
2. Which country offices run hot (Hot offices can lead to employee complaints and reduced productivity) and which offices run cold (Cold offices cause elevated energy expenditures and employee discomfort). Calculate count of offices run in hot and count of office run in cold by country.
3. Our data set includes information about the performance of five brands of HVAC equipment, distributed across many types of buildings in a wide variety of climates. We can use this data to assess the relative reliability of the different HVAC models(i.e We can see that the which model seems to regulate temperature most reliably and maintain the appropriate temperature range). Calculate count of extreamtemp by hvacproduct

Contact Us

Visit us on: http://www.analytixlabs.in/

For more information, please contact us: http://www.analytixlabs.co.in/contact-us/

Or email: info@analytixlabs.co.in

Call us we would love to speak with you: (+91) 95-55-219007

Join us on:

Twitter - http://twitter.com/#!/AnalytixLabs

Facebook - http://www.facebook.com/analytixlabs

LinkedIn - http://www.linkedin.com/in/analytixlabs

Blog - http://www.analytixlabs.co.in/category/blog/