Data Analyst Takehome

Create Cohort Repayment Curves

Introduction

This take-home exercise is your opportunity to demonstrate your Data Analyst skills!

You will be given some fake data that is representative of the kind of data we work with a M-KOPA and be asked to transform that data in order to answer a business question. This test is designed to assess the following:

- Ability to interpret and understand a business analysis request
- SQL skills
- Data interpretation and analysis presentation skills

We will use your take-home response as a basis for the technical interview and explore some of the finer details about how your proposal could be used.

Housekeeping:

- This take-home should take around 3-4 hours
- Please return submissions to natalie.mbuvi@m-kopa.com
- If you have any questions, please email the address above.
- We expect submissions within 1 week of receiving the task

M-KOPA's **customers** open a pay-as-you-go **account** for a specific **plan** and make **payments** over time to purchase a product. This process is represented by the following datasets provided to you in separate csvs.

Account

	AccountId	RegistrationDate	CustomerId	PlanId
1	5001	2020-01-19 00:00:00.0000000	2127	66
2	5005	2020-07-01 00:00:00.0000000	3117	55
3	5006	2020-03-19 00:00:00.0000000	2383	44
4	5007	2020-10-06 00:00:00.0000000	3664	51
5	5012	2020-09-29 00:00:00.0000000	3604	49
6	5015	2020-06-22 00:00:00.0000000	1115	41
7	5017	2020-08-18 00:00:00.0000000	3212	68

Customer

	CustomerId	FirstName	LastName	Region
1	1000	Frank	Chocho	kisumu
2	1003	Victor	Githui	nairobi
3	1004	Angela	Lwoyelo	nairobi
4	1007	Obinna	Muguku	kisumu
5	1008	Dorcas	Lwoyelo	nairobi
6	1009	Mercy	Muguku	kisumu
7	1012	Dorcas	Lwoyelo	kisumu
8	1013	Mercy	Nyakwea	nairobi
9	1014	Titus	Warimu	kisumu

Payment

	PaymentId	Amount	ReceivedWhen	AccountId	PaymentType
1	1006	160.78216374269	2021-04-11 00:00:00.000	6737	DailyPayment
2	1008	97.8991228070175	2020-09-12 00:00:00.000	6263	DailyPayment
3	1009	43.947585978836	2020-08-27 00:00:00.000	8111	DailyPayment
4	1010	36.6279761904762	2020-06-01 00:00:00.000	8194	DailyPayment
5	1011	155.324115044248	2020-07-10 00:00:00.000	7813	DailyPayment
6	1013	129.125	2021-02-13 00:00:00.000	6875	DailyPayment
7	1014	155.077160493827	2020-08-03 00:00:00.000	5487	DailyPayment
8	1015	104.131433823529	2021-01-10 00:00:00.000	6067	DailyPayment

PaymentPlan

	PaymentPlanId	Product	DailyValue	LoanTerm	Deposit	TotalValue
1	10	phone	25	200	250	5250
2	11	solar	50	200	1000	11000
3	15	tv	65	300	975	20475
4	16	tv	65	300	975	20475
5	17	tv	55	200	1100	12100
6	18	tv	65	200	975	13975
7	19	phone	35	150	175	5425
8	20	solar	45	200	675	9675

Context

The Credit team want you to create cohort repayment curves for M-KOPA's portfolio.

A cohort is a subset of customers who were all registered in the same month. A Cohort Repayment Curve is the cumulative percentage paid of the total cohort value at each month since registration (months on books). They want to filter these cohort curves by Product and Region.

A Cohort Repayment Curves show the percentage paid of the cohort TotalValue by the months on books.

$$PercentagePaid_{i,j} = \frac{\sum_{k=0}^{K} AmountPaid_{j,k}}{\sum_{k=0}^{K} TotalValue_k}$$
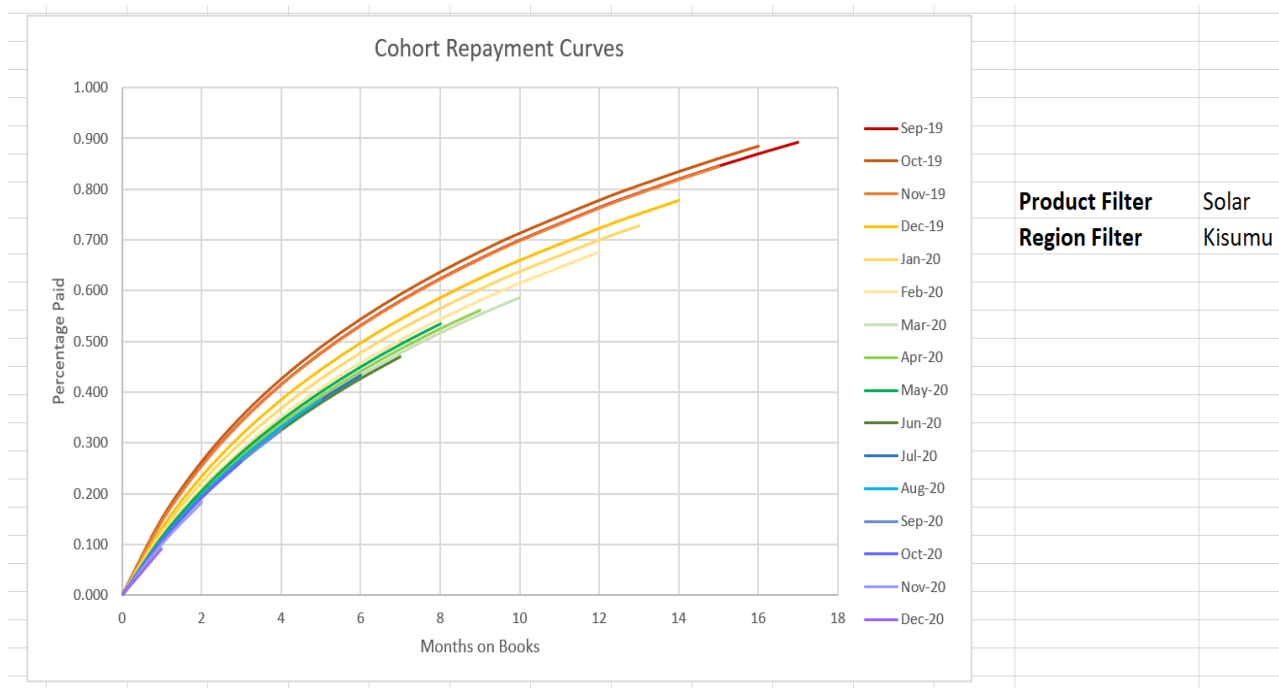
i = the cohort
j = the months on book
k = all accounts registered during cohort i

AmountPaid$_{j,k}$ = the amount paid by account k by month on book j
TotalValue$_k$ = the total value to be paid by account k

Example Repayment Curves – Note the repayment curves you generate are not expected to look exactly like this

Task Description

Your task is to use SQL to transform the raw data provided to create a dataset that allows you to generate the repayment curves requested. Please provide material for the SQL transformation and any downstream analysis of the resulting dataset in the tool of your choice (excel, python, R etc)

Once the repayment curves are created, please provide a summary document where you:

- Discuss any interesting nuances in the raw data

- Any decisions or assumptions you made during the SQL transformation

- A discussion of any patterns or insights you can derive from the final transformed dataset and visual.

Your response should :

- Include all analysis code (SQL, Excel, python, R etc)
- A summary document

Stretch goals:

- Explore the raw payment data – what insights can you derive about payments from that?
- Can you do the same data transformation in python or R? What are the tradeoffs between using different approaches?