**HMDA Data Project**

**Problem Statement and Instructions**

**Problem Statement**

Change Financial is a small regional bank headquartered in Washington, DC, and has branches in surrounding states. It’s time to grow, and Change’s leaders want to know if entering the home loans market is a smart strategic move. Being an information-driven culture, they first want to understand their region’s home lending market.

Your partners on the Change strategy team have pointed your team towards the Home Mortgage Disclosure Act (HMDA) data set collected and published each year by the federal government and which has basic data on all home loan applications received by lenders across the U.S. The data is made available for analysis in support of public policy ([HMDA History)](https://www.ffiec.gov/hmda/history.htm), and is a valuable as a source of information about the market.

The data team’s idea is to build a data product to help the strategists answer questions about the regional home loans market; in particular, to help understand both the “Conventional/Conforming” loans market. You’ve volunteered to create the first working iteration so the team can begin planning its workweeks. After meeting with the strategy team and your team, you’ve got an idea of where to start, key concerns, and how you can help your team work with the market data.

**Instructions**

As you start the challenge, realize that this is real-world, imperfect data. If you find yourself uncertain of what the “right” answer is, use your best judgment, make an assumption, document it, and keep going.

Overall, we first ask you to show your data skills in three areas at a basic level, and then, in the last step, you’ve got the freedom to take this data product in your own direction.

1. **Data munging** – get the data loaded and then provide a data export function. Keep in

mind, your team members will use this to get started when they work with this data

a. Merge the application data with institution data such that each loan application is assigned a “Respondent\_Name”. Additionally, create a new attribute that buckets

“Loan\_Amount\_000” into reasonable groups b. Provide a simple API of two functions:

hmda\_init() – Read the data files and return a pointer or object containing the expanded HMDA data from part a.

hmda\_to\_json(data, states, conventional\_conforming) – Export the expanded data set to disk for the states filtered by product segment. Both

parameters should be optional, and you can add more optional parameters of

your own if you’d like

2. **Quality check** – bad data is worse than no data at all, and both customers and peers will need to understand what you learn here

a. Perform quality assessment on those *Loan\_Amount\_000* and

*Respondent\_Name*. Go beyond simple descriptive statistics; provide rules or techniques for quality monitoring that can be applied each year the data gets updated.

b. Explain an additional column or two that you think is important to monitor and why

c. Document in your metadata any concerns on data quality as well as recommendations for discussion that you could present in-person

3. **Craft a visual data narrative** – imagine you have 20 minutes to update the strategy

team on what the data has and what you’ve learned. You need some visuals to illustrate some insights. *Charts and plots must be generated from your code; not from produced in external standalone software like Excel or Tableau*

a. Provide a hypothesis about Change Financial’s decision to enter a new geography

b. Visualize metrics for market size by state and year; provide a business

explanation for what the chart shows or explains and whether your hypothesis is proved, disproved, or you need additional data

c. Summarize your key insights based on the visualizations

4. **Finished steps 1 – 3? Take it a step further!** What piqued your curiosity? Are you worried about data quality issues? Do you think it’s too hard to get market insights? Hypothesize a business need and prototype a tool of your choosing to improve this data

product and highlight your area of expertise. To prime your idea pump, here are some

“thought starters”:

*What if quality was interactive?* Provide an interactive or systematic tool to clean outliers, *e.g.*, census tract-by-tract checking for all tracts, way and extend your code with functionality to fix them. Automatically update metadata

*Another direction, what if we went the analytic route?* Identify clusters of lending patterns, *e.g.*, geographies, that represent markets where we want to market

similar products

*Or, what if we had deep historical data?* Add historical HMDA data (the metadata has tips on this) such that we can see trends for a region. Offer choices for what kinds of trends to see and where (*e.g.*, a volume trends tool would allow us to

see trends across a choice of states)

*Idea #4, interaction is always cool, can it work here?* Create an interactive experience that is designed to allow your VP to see market share for a competitor in a given geography

**Data and Tools**

*Solutions that require purchase of a software license, agreement to a free trial period, or purchased access to data will not be accepted regardless of whether or not Capital One uses said software or data. Some examples of licensed software that are not acceptable for this challenge include SAS, Tableau, and Excel. Abide by all applicable laws and regulations regarding the use of software or external data sources. If you have questions about a particular software package, please contact your recruiter immediately.*

**Data**

**Downloading the data is a simple two-step process: Please use the same web browser for both links.**

**1. Please access the Capital One Data Challenge GitHub account via** [**https://github.com/login**](https://github.com/login) **using the Username and Password provided in the email from your recruiter.**

**2. Once logged in to GitHub, please copy and paste the following link into your web browser and press enter to download the ZIP file.**

[**https://github.com/c1-data-analytics/data-challenge-data/archive/master.zip**](https://github.com/c1-data-analytics/data-challenge-data/archive/master.zip)

**The ZIP file will contain the following list of data and metadata files necessary for you to work through this Data Challenge. Please do not change the username or password while accessing this account.**

|  |  |
| --- | --- |
| **Resource** | **You should see** |
| This document | HMDA analysis – Data Challenge.docx |
| Technical Considerations | Data\_Challenge\_Technical\_Considerations.html |
| Main input data sets | 2012\_to\_2014\_loans\_data.zip  2012\_to\_2014\_institutions\_data.zip |
| Additional metadata for the data | Data\_Challenge\_Metadata.html |
| Optional HMDA metadata | 2013guide.pdf, in particular Appendix A |

**Tools**

We want to see how you solve a business problem using programming tools, as such, spreadsheet-based analysis or munging is not allowed. Here are some example platforms you should feel free to use. You are not limited to this list, and our solution review team will be able to evaluate solutions in most languages. If you really do have a question about the platform you would like to use to solve the problem, contact your recruiter with the exact setup you’d like to use (including OS and specific versions when applicable), your backup choice, and they can seek verification for the platform

|  |  |
| --- | --- |
| **Platform example** | **Notable packages** |
| Anaconda Python Distribution | notebook, pandas, matplotlib, bokeh |
| R | R, Shiny, plyr, ggplot |
| JavaScript | D3, nvd3, node.js |
| Java virtual machine | Groovy, Scala |
| Other software packages with which you are familiar | |