# Project 1

**To Do:**

1. Choose a dataset [Kaggle.com](http://Kaggle.com)
2. Project Proposal - Due on Thursday 05/30/24
3. Guiding questions - 3 Hypotesis questions
4. inspiration
5. Example visualization

**Resources:**

<https://bootcampspot.instructure.com/courses/6446/pages/7-project-1-overview?module_item_id=1248939>

coolors.co

Rubric:

I don't care what canvas thinks about Project 1

what is your dataset and why

three research guiding questions

inspiration - other code/articles similar to your dataset

example visualization? you need at least 5 - typically 1-2 per research question

what are you regressing?

color palette

roles & responsibilities

who is answering which question and who is doing the regression

all do the data cleaning and all do presentation prep

partner and group coding is encouraged

are we presenting this like a powerpoint, or just presenting our code in jupyter?

and GitHub link

EVERYTHING FOR THE PROPOSAL

## **FINAL DELIVERABLES**

Presentation JUNE 10

Canva

powerpoint or google slides

“*canva is like pretty powerpoint”*

## **Presentation**

15-20 min  
EVERYONE must speak

intro the project

go over the data cleaning

null values, datatypes, new columns, aggregations

1-2 slides max

3-4 data story chapters 1 - 2 slides each

regression - 1 slide or 2

conclusions - CALL TO ACTION

data driven decision.

15min of presentation - 1 min or 2 for Q&A  
The Main Room - 1 Breakroom  
9 projects @ 20 min takes about 3 hours

LIMITATION AND FUTURE WORK

Works Cited - link to your data, GitHub, articles or inspiration, whatever

q&a, thank you

business casual  
professional presentation geared to a mostly technical data science project lead  
1 notebook for data cleaning  
3-4-5 notebooks - 1 per research question and regression  
Slide deck  
project proposal  
 Writeup  
IS NOT DUE until June 17  
clean up the GitHub, clean up the writeup.

**Others:**

you can add more details, appendixes, or lots of vizzes  
nuances of the data cleaning, of each of the questions  
calendar:  
project proposal is due by Thursday  
data cleaning should be mostly done or about ready on by mid class on Monday  
Tuesday is hard core async question and regression work  
Thursday coming back and putting together the slides  
presentation Monday.

A screenshot of a computer

Description automatically generated

**git kraken**  
search for git kraken  
git kraken can be used by ANYONE  
but hopefully at least 1 project owner  
OTHER UIs = GitHub desktop, or a terminal like Git Bash  
THIS IS OPTIONAL, but recommended for projects, I like it at least

1. downloaded GitKraken
2. Linked it to our personal github

ALWAYS FREE - you can this for your public repos

INDIVIDIUAL contributions (not projects)

.DS\_Store

1. ONE PERSON PER GROUP  - edit the .gitignore

# STANDUPS

# STANDUPS

ONCE HERE AND ONCE IN [#18-project1](https://datapteastapr-h6t4255.slack.com/archives/C075EL4DK6H)

[8:27](https://datapteastapr-h6t4255.slack.com/archives/C06PPQCRQLS/p1716946070747939)

Alexander Booth (Instructor)

1. WHAT DID YOU DO TODAY

[8:28](https://datapteastapr-h6t4255.slack.com/archives/C06PPQCRQLS/p1716946096226759)

Alexander Booth (Instructor)

2) WHAT IS THE PLAN FOR TOMORROW (AKA a day off OR next class)

[8:28](https://datapteastapr-h6t4255.slack.com/archives/C06PPQCRQLS/p1716946113667429)

Alexander Booth (Instructor)

A NEAR-TERM FUTURE GOAL (edited)

[8:29](https://datapteastapr-h6t4255.slack.com/archives/C06PPQCRQLS/p1716946143153269)

Alexander Booth (Instructor)

3) What did you run in trouble with (aka what is/was a BARRIER)

[8:29](https://datapteastapr-h6t4255.slack.com/archives/C06PPQCRQLS/p1716946163989529)

Alexander Booth (Instructor)

WE HAD NO BARRIERS TODAY

1. WHAT DID YOU DO TODAY
   1. Installed GitKaken
   2. Choose Dataset
2. 2) WHAT IS THE PLAN FOR TOMORROW (AKA a day off OR next class)
3. 1. Determined question o hypothesis
4. Define resources.
5. Distribute task (Cleaning data, inspiring arguments, etc)

A screenshot of a computer

Description automatically generated

so you may want to look at what other types of visualizations you can do on categorical data yeah you got a lot of flags you can have a lot of Catholic names so you have a decent number of numeric data sets but I think it's probably fine the only thing other thing you guys have to look out to you all on how much work you want to do this is a if you're going to do a predictive model you cannot do a linear regression on what's the column I think it's flat what's the what's the column for whatever they get approved o

oh yeah yeah I think it's on the credit report this data set so I don't think y'all feel too strongly about it so I would probably find it is better data sets to use if you want loan predictions you can just Google alone predictions if you want but I would probably feature this data set if and I do have rights so you would have to convince me very strongly that you want to use the states

I am I got a big favor to ask of you yes so this is a group of five and I think we misjudged and created a group of three so would you be open and migrating to a different project no one's really started anything yet so yeah let me get that straightened up here and then I will introduce you to your neighbor and I'm sorry for that missed calculation here we had a yeah we're having a side chat about things in the T

Grupo 6

WHAT DID YOU DO TODAY

Installed GitKaken

Choose Dataset

WHAT IS THE PLAN FOR TOMORROW (AKA a day off OR next class)

Determine question or hypothesis.

Define resources.

Distribute task (Cleaning data, inspiring arguments, etc.)

Solved problem with Installed Gakken

Resourses

https://github.com/umairinayat

## **Resources about dataset**

Inspiring

<https://www.kaggle.com/code/mohamedkhaledmahmoud/loan-eligibility-prediction-model>

<https://www.kaggle.com/code/umairinayat/loan-eligibility-with-various-ml-models>

<https://github.com/hmcninson/loaneligibility>

<https://github.com/limchiahooi/loan-approval-prediction>

https://www.youtube.com/watch?v=T9kgWBmUIRk&t=17s

https://github.com/Solution92/Loan-Approval-Prediction-Analysis-with-Power-BI

|  |  |  |
| --- | --- | --- |
| Item | What | url |
| 1 | In the lending industry, investors provide loans to borrowers in exchange for the promise of repayment with interest. If the borrower repays the loan, then the lender would make profit from the interest. However, if the borrower fails to repay the loan, then the lender loses money. Therefore, **lenders face the problem of predicting the risk of a borrower being unable to repay a loan**. | https://www.kaggle.com/code/ridhijhamb/loan-prediction-random-forest |
| 2 | to predict loan eligibility and streamline Dream Housing Finance Company's decision-making process. |  |
| 3 | Dream Housing Finance company specializes in providing all types of loans across urban, semi-urban, and rural areas. Upon receiving a loan application, the company undertakes a validation process to determine the customer's eligibility. This validation process is based on various customer details provided during the online application, including age, house ownership, working experience, intent of loan, income, loan amount, credit history, among others. To streamline and automate this loan eligibility process in real-time, the company aims to identify specific customer segments that qualify for loan amounts. By targeting these eligible customer segments, the company can optimize its loan offerings and better serve its clientele. | https://github.com/Akshat8303/Loan-Prediction/tree/main |
| 4 | The project will also focus on developing predictive models to assess loan default risks and borrower creditworthiness. | https://github.com/Sau2901/Loan-data-set-analysis |
| 5 | Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan. | https://github.com/shrikant-temburwar/Loan-Prediction-Dataset |
| 6 | This project focuses on **predicting loan approval outcomes** through an extensive analysis of a curated dataset | https://github.com/ronbodnar/loan-eligibility-prediction/tree/main |

My propuse

Which one is the factor impact in the making decision to loan approval or rejection.

Demographic factors have a higher impact making decision loan approval.

Which one is the association or correlation between loan amount and repayment history.

The more dependents and married is likes to be approved loan.

Null Hypothesis (H₀)

**Applicants with more dependents and those who are married tend to have a higher likelihood of loan approval.**

**Alternative Hypothesis (H₁)**

**Applicants with more dependents and those who are married tend to have a lower likelihood of loan approval.**

**Question**

**What factors influence the decision for loan approval?**

**Question**

Which demographic factors have the most influence on loan approval decisions?

**Jobs and Salaries in Data Science**

**https://www.kaggle.com/code/alfathterry/hypothesis-testing**

<https://www.kaggle.com/code/ziadmostafa1/data-jobs-analysis>

<https://www.kaggle.com/code/umeshlohakare/data-analysis-data-science-job-and-salary>

<https://www.kaggle.com/code/ramaamayri/exploring-job-trends-and-salaries-in-data-felid>

**https://www.kaggle.com/code/fahadrehman07/data-science-job-salary-prediction-glassdoor/notebook**

analysis

<https://github.com/TomaIjatomi/EDA-on-Data-Science-Job-Salaries/blob/master/DS_Salaries.ipynb>

<https://github.com/ALOTAIBIH/Data-Science-Jobs-and-Salaries-Analysis/blob/main/project_2_DATA_science%20job%20analusis.ipynb>

<https://github.com/luuisotorres/Data-Science-Job-Salaries-EDA-and-Predictions-with-PyCaret/blob/main/job-salaries-eda-and-predictions.ipynb>

<https://github.com/Subhralina/Data-Science-Salary-Prediction/blob/main/exploratory_data_analysis.ipynb>

https://github.com/agneszhrn/Data-Science-Job-Salaries-Analysis-and-Predict

Difference between regression and correlation

Analysis and interpretation. Discussion

|  |  |
| --- | --- |
| Question 3 | Looking at attrited customer, identify common characteristics, and make suggestions on how to retain customers   1. Box plot 2. Bar chart |

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated