Course Projects

Assessment Schedule

Submission 1

- Project Proposal (5%)
- Descriptive Statistics and Exploratory Data Analysis (EDA) (5%)

Submission 2

- Data Cleaning/Transformation (5%)
- Data Analysis (10%)

Submission 3

- Data Product and Team Presentation (15%)

Random Course Project Tips

- Student IDs are confidential, don't list it on your slides
- Your picture on the slides introducing your team is a nice personal touch.
- Visualizations and plots should be clear.
  - Label your X and Y axis, label categories, etc.
- Proper Citation
  - I would rather receive your less than perfect English than stolen English.
  - Properly cite your work.
  - You can include code from others if you properly references it and acknowledge the source.
- Don't forget about the "science" in "data science".
  - Background/Motivation
  - Define the problem/issue/pattern you are investigating
  - Relevance/Significance
  - Research Questions
  - Can the questions be answered from the data you have?
- Class code example demos are not a script to be blindly followed on your project. Use your good judgement. Think about what you delivering.

Submission Instructions

- Submit presentation slides, dataset link and code.
- The slides should target a business audience (do not include code as part of your slides).
- For submission 1 (problem statement and EDA), limit your submission to defining the problem, motivating the problem and understanding the data you will be working with. Don't submit any analysis that answers the questions from the problem statement.
- You are expected to use Python for data cleaning and pre-processing
- Code must be posted on a public code repository and uploaded on Blackboard

Project Proposal

- What is the problem you are trying to solve or insight you are looking to gain with the support of data?
- What is the background and motivation for addressing the problem?
  - Why is the project important?
- Who benefits from this project?
  - Who is the target audience for your analysis?
- What are the questions you are looking to answer or objectives you are looking to achieve from this project?
- You can adopt one of two approaches:
  - Descriptive Analysis
  - Prediction Problem

Exploratory Data Analysis (EDA)

- Describe the data
- Does the data include missing, incomplete or invalid records?
- Does your data include outliers?
- Is the data segmented in groups?
- Is the data imbalanced (large number of the records represent a majority class and very few records represent the minority class)?
- Are some data elements highly correlated with each other?
- How was the data collected?
- What are the inclusion criteria for your data?
- Can you generate preliminary visualizations for individual features?

Project Tasks (1 of 4)

1) Business Problem and Dataset
   - Confirm with your instructor prior to submission
   - Use the slides to understand the scope
2) Development Environment
   - Select technology stack, libraries, code repository, document sharing, etc.
3) Load Data
   - Load into your development environment

Project Tasks (2 of 4)

4) Exploratory Data Analysis (EDA)
   - Confirm the data is correctly loaded
   - Identify data types (categorical, numerical)
   - Check validity of data
   - Define schema
   - Understand the data
   - Answer questions from EDA slides
   - Use visualization to understand and explore not to explain

Project Tasks (3 of 4)

5) Data Cleaning/Transformation
   - Resolve issues identified during the EDA phase
     - Missing, incomplete, duplicate, etc.
   - Data Transformation
     - Merge data into a single DataFrame
     - Rename columns
   - Save a cleaned up/transformed copy of your dataset to disk, so that you don't need to repeat those steps during the analysis phase

Project Tasks (4 of 4)

6) Data Analysis
   - Identify key patterns
   - Use visualization to explain
   - Answering the questions from the problem statement
7) Data Product and Team Presentation
   - 10-minute recorded presentation

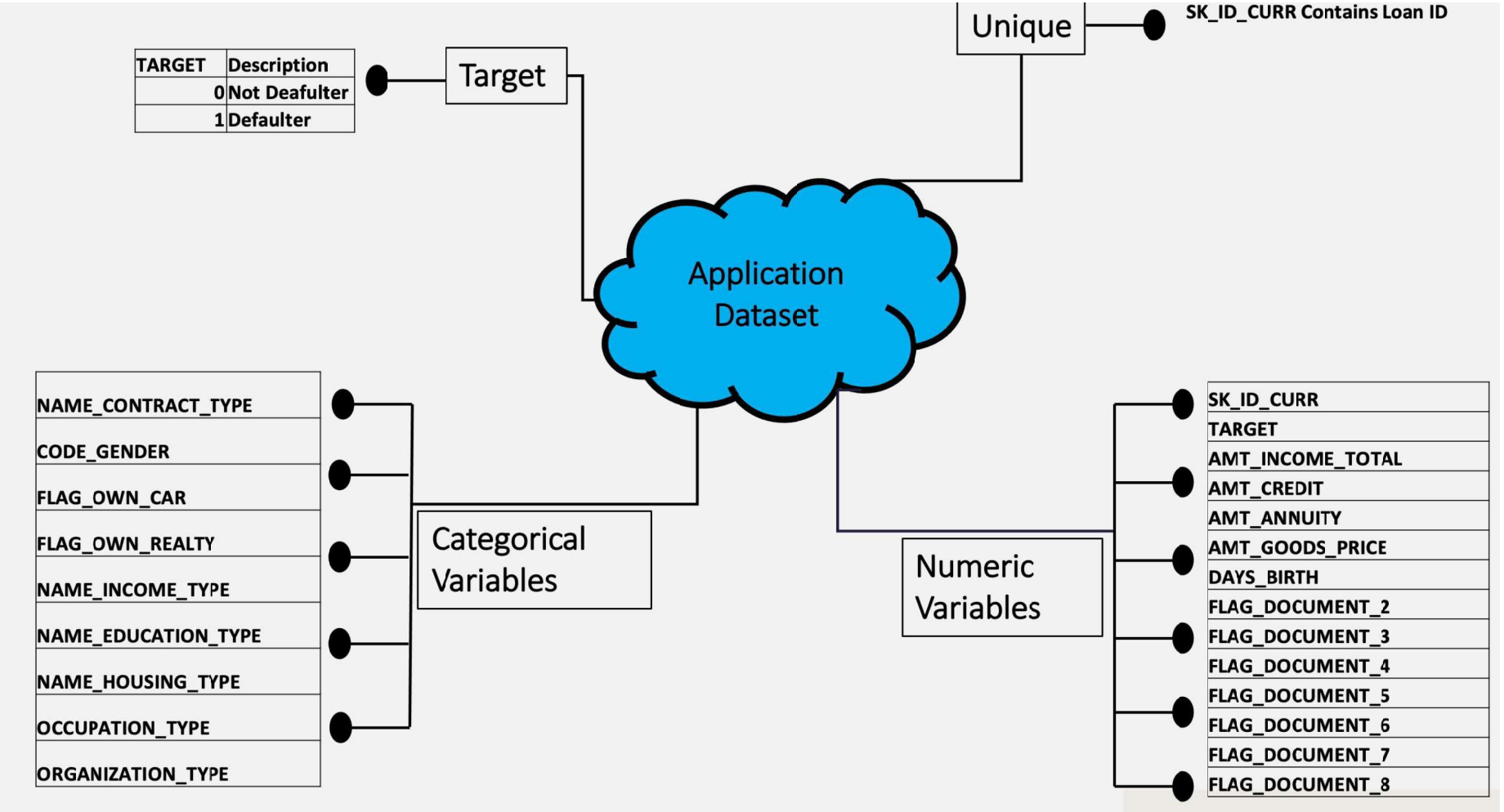
Project Proposal Fake Example

- **Team Introduction**
  - Include Group Number, Section Number
  - Include Individual Team members
- **Background/Motivation**
  - The App Store has over 2 million apps and over 3 thousands new Apps approved very hour. It is an increasingly complex and crowded space which is difficult for new developers to get noticed.
- **Problem Statement**
  - Application developers need better insight into what application categories are popular on the App store in order to target their development effort towards areas in high demand and maximize their profit potential
- **Project Proposal**
  - Our project team will create a descriptive analytics product to examine trends in app store downloads to help application developers gain insights into marketplace trends that can be useful in developing popular applications.

Project Proposal Fake Example

- **Analysis Questions**
  - What are the top categories of downloaded apps?
  - What are the most popular application categories by age group?
  - Which apps lead to a higher conversion rate from free to premium version?
- **Dataset Description**
  - The analysis will be based on XYZ dataset obtained from ABC source.
  - The dataset included data between year 1 and year 5
  - Include analysis results form EDA describing the dataset and including preliminary visualization
  - Explain transformation steps do you expect to perform on your dataset prior to processing

Data Description - Good



No Code on Slides Please

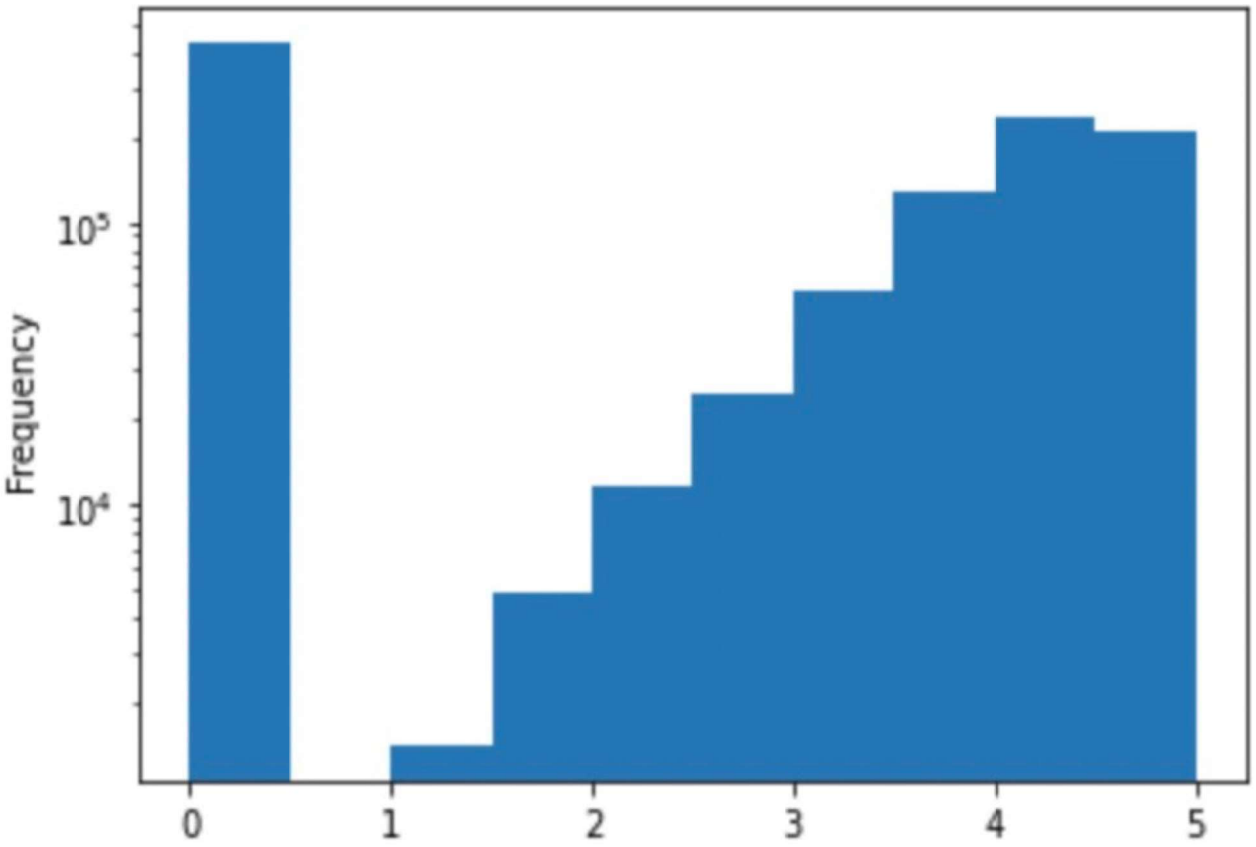
```
type(gplst)
```

```
pandas.core.frame.DataFrame
```

```
gplst.shape
```

```
(1118136, 23)
```

No Missing Plot Labels Please



Need to

- Label X-axis and Y-axis
- Include Title

Project Proposal Grading Rubric

Assessment Criteria	Below Expectations 0	Good 0.5	Very Good 1.0
Adherence to Submission Instructions	Did not follow instructions	Some mistakes or nor clearly delivered	Followed all project submission instructions
Background/ Motivation	Missing or poorly presented	Some mistakes or nor clearly delivered	Clearly explained and presented and free from errors
Problem Statement	Missing or poorly presented	Some mistakes or nor clearly delivered	Clear problem statement identifying the need for the project and intended audience. Demonstrate clear understanding of the problem
Project Proposal	Missing or poorly presented	Some mistakes or nor clearly delivered	Clearly defined and well written project objectives
Analysis Questions	Missing or poorly presented	Some mistakes or nor clearly delivered	Questions related to the project objectives, can be directly answered from the data, and add valuable insight

EDA Grading Rubric

Assessment Criteria	Below Expectations 0	Good 0.5	Very Good 1.0
Adherence to Submission Instructions	Did not follow instructions	Some mistakes or not clearly delivered	Followed all project submission instructions.
Data Source Description	Missing or poorly presented	Some mistakes or not clearly delivered	Data Source clearly explained, inclusion criteria, collection method.
Descriptive Statistics	Missing or poorly presented	Some mistakes or not clearly delivered	Statistical methods are fully and correctly applied.
Exploratory Data Analysis (EDA)	Missing or poorly presented	Some mistakes or not clearly delivered	Data are completely and appropriately interpreted. Identified groups, patterns, anomalies, outliers.
EDA Visualization	Missing or poorly presented	Some mistakes or not clearly delivered	Visualizations are correctly constructed and clearly demonstrate data interrogation.