[](http://localhost:8888/tree)

Logout

project1-starter Last Checkpoint: a few seconds ago (autosaved)



Python 3

Trusted

* File
* Edit
* View
* Insert
* Cell
* Kernel
* Widgets
* Help

**Project 1¶**

In this first project you will create a framework to scope out data science projects. This framework will provide you with a guide to develop a well-articulated problem statement and analysis plan that will be robust and reproducible.

**Read and evaluate the following problem statement:¶**

Determine which free-tier customers will covert to paying customers, using demographic data collected at signup (age, gender, location, and profession) and customer usage data (days since last log in, and activity score 1 = active user, 0= inactive user) based on Hooli data from Jan-Apr 2015.

**1. What is the outcome?¶**

Answer: The outcome would be to determine free-tier customer(s) that might become paying customer(s).

**2. What are the predictors/covariates?¶**

Answer: The predictors/covariates are as follows: age; gender; location; profession; last login; score

**3. What timeframe is this data relevant for?¶**

Answer: This data is relevant for Jan-April 2015 timeframe

**4. What is the hypothesis?¶**

Answer: Relationship exists indicating connection between the demographics of a customer, their use of the product and the high possibility that they may or may not not decide to start paying for the service.

**Let's get started with our dataset¶**

**1. Create a data dictionary¶**

Answer:

​

Variable | Description | Type of Variable

---| ---| ---

Var 1 | 0 = not thing 1 = thing | categorical

Var 2 | thing in unit X | continuous

​

We would like to explore the association between X and Y

**2. What is the outcome?¶**

to the

Answer: Whether or not a student was admitted to the UCLA Graduate School

**3. What are the predictors/covariates?¶**

Answer: The GRE Score and also the Grade Point Average attained from the previous school attended

**4. What timeframe is this data relevant for?¶**

Answer: The timeframe was not provided

**4. What is the hypothesis?¶**

Answer: UCLA Grad School admission decision is focused on a student's GRE grades, Grade Point Average, and the student's previuos school's ranking.

Using the above information, write a well-formed problem statement.

**Problem Statement¶**

**Exploratory Analysis Plan¶**

Using the lab from a class as a guide, create an exploratory analysis plan.

**1. What are the goals of the exploratory analysis?¶**

Answer:

**2a. What are the assumptions of the distribution of data?¶**

Answer:

**2b. How will determine the distribution of your data?¶**

Answer:

**3a. How might outliers impact your analysis?¶**

Answer:

**3b. How will you test for outliers?¶**

Answer:

**4a. What is colinearity?¶**

Answer:

**4b. How will you test for colinearity?¶**

Answer:

**5. What is your exploratory analysis plan?¶**

Using the above information, write an exploratory analysis plan that would allow you or a colleague to reproduce your analysis 1 year from now.

Answer:

**Bonus Questions:¶**

1. Outline your analysis method for predicting your outcome
2. Write an alternative problem statement for your dataset
3. Articulate the assumptions and risks of the alternative model

In [ ]:

​