Project 1: Solutions

Read and evaluate the following problem statement:

Determine how likely free-tier customers are to convert to paying customers, using demographic data collected at signup (age, gender, location, and profession) and customer useage 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: Likeliness of conversion to paid customer

2. What are the predictors/covariates?

Answer: age, gender, location, profession, days since last log in, and activity score 1 = active user, 0= inactive user

3. What timeframe is this data relevent for?

Answer: Jan-Apr 2015

4. What is the hypothesis?

Answer: That customers who were more recently active are more likely to convert to the paid teir

Let's get started with our dataset

In [1]:

```
%matplotlib inline
import matplotlib.pyplot as plt
import pandas as pd
import statsmodels.api as sm
import pylab as pl
import numpy as np

df = pd.read_csv("../assets/admissions.csv")
df.head()
```

Out[1]:

	admit	gre	gpa	prestige
0	0	380	3.61	3
1	1	660	3.67	3
2	1	800	4.00	1
3	1	640	3.19	4
4	0	520	2.93	4

1. Create a data dictionary

Answer: To be completed when the dataset is finalized

Variable	Description	Type of Variable
Admit	0 = not admitted 1 = admitted	categorical
GRE	GRE score 200-800	continuous
GPA	GPA 0-4.0	continuous
Prestige	1= not prestigious 2 = low prestige 3= good prestige 4= high prestige	categorical

We would like to explore the association between admission into grad school and the prestige of undergraduate institutions.

1. What is the outcome?

Answer: admission into grad school

2. What are the predictors/covariates?

Answer: Prestige, GRE, GPA

3. What timeframe is this data relevent for?

Answer: The timeframe for this data isn't immediately clear. This is something that could be researched further by contacting the original collectors of the data or researching the dataset's history. It could also be acknowledged as a potential limitation of the dataset.

4. What is the hypothesis?

Answer: Students that more prestigious undergraduate schools will have higher admissions rates into graduate school.

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

Problem Statement

Determine if there is an association between graduate school admission and the prestige of a student's undergraduate school using data from the UCLA admissions data set.

Exploratory Analysis Plan

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

1. What are the goals of the exploratory analysis?

Answer:

1. Determine if there is any missing data
2. Examine the distributions of the variables to determine if any of the variables need be transformed

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

Answer: normality

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

Answer: histograms

3a. How might outliers impact your analysis?

Answer: They could skew the associations in the direction of the outlier.

3b. How will you test for outliers?

Answer: Box plots are one good way.

4a. What is colinearity?

Answer: when two variables are capturing similar variance in the data

4b. How will you test for colinearity?

Answer: create a correlation matrix

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:

1. Check for missing data and remove observations.
2. Check for colinearity.
3. Check for normal distribution.

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 [ ]:
```