DSCI 101 Sample Final Exam

Instruction

- This exam is a timed exam. You have 2 hours to complet it.
- You may access any course material including slides, demo, notes and homework.
- You may access other related books or reference material you can find.
- You may also use the internet in a **passive way**. That is, you may search on the internet for material, but you **cannot** post a question and ask for help, including asking ChatGPT or other AI tool for help.
- If you do use something significant from the internet, please put the link to the source in your anwser. For example: you copy some formula or take a block of code directly from some website.
- You **cannot** get help from anyone or discuss exam problems with **anyone inside or outside the class**.
- If you need clarification on anything, please ask the instructor directly.
- Feel free to include any Python modules you need besides the ones included already.
- Feel free to add more cells for codes and texts, and please make your code sufficiently commented and readable.

Race and Policing in San Francisco, CA (https://en.wikipedia.org/wiki/San_Francisco)

Data we are using in this exam is from the Stanford Open Policing Project (https://openpolicing.stanford.edu/data/), a national repository of traffic stop and search data, to examine racial disparities in policing. We will look at the city of San Francisico.

In [ ]:
```
### standard imports
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
# ignores warning message
import warnings
warnings.filterwarnings('ignore')

# allows you to view the plots upon executing your code
%matplotlib inline
# sets the plotting style, feel free to change!
plt.style.use('fivethirtyeight')
# prevents histogram bars to fuse together
plt.rcParams['patch.force_edgecolor'] = True

# fix random seed for reproducibility
np.random.seed(2024)
```

1. Import

Read in the data from the following url directly, save as a dataframe named `police_stop` . Then anwser **Question 1 - 5** on Canvas. **DO NOT download the file and read in from your local drive**. When I run your code, I will not have the data file in my local drive.

- **data file url:**
  - `https://stacks.stanford.edu/file/druid:yg821jf8611/yg821jf8611_ca_san_francisco_2020_04_01.csv.zip`

- Question 1: How many rows in the dataframe `police_stop` ?
  - A. 905050
  - B. 905070
  - C. 907050
  - D. 907070
- Question 2: How many columns in the dataframe `police_stop` ?
  - A. 20
  - B. 21
  - C. 22
  - D. 23
- Question 3: Which of the following is NOT a column in the dataframe `police_stop` ?
  - A. date
  - B. time
  - C. search_conducted
  - D. raw_ethnicity
- Question 4: Which of the following columns does NOT contain missing values?
  - A. subject_age
  - B. contraband_found
  - C. reason_for_stop
  - D. search_vehicle
- Question 5: Which of the following is correct about the the number of columns in the dataframe `police_stop` that has the following particular data type?
  - A. integer, 1
  - B. float number, 3
  - C. bool, 4
  - D. object, 15

In [ ]:
```
### your code here
```

2. Clean

- Drop all the columns start with `raw_` . Your dataframe after dropping should still be named `police_stop` . Then anwser **Question 6 - 10** on Canvas.

- Question 6: How many categories (distinct values) are there in the column `subject_race` ?
  - A. 4
  - B. 5
  - C. 6
  - D. 7
- Question 7: What is the earliest and latest date in the dataframe `police_stop` ?
  - A. Jan 1st, 2007, Dec 31, 2016
  - B. Jan 1st, 2008, Dec 31, 2016
  - C. Jan 1st, 2007, Jun 30, 2016
  - D. Jan 1st, 2008, Jun 30, 2016
- Question 8: Which year has the most number of stops recorded in the dataframe `police_stop` and how many?
  - A. 2007, 113125 stops
  - B. 2008, 113125 stops
  - C. 2009, 116012 stops
  - D. 2010, 116012 stops
- Question 9: Which day of the week has the most number of stops recorded in the dataframe `police_stop` ?
  - A. Monday
  - B. Tuesay
  - C. Wednesday
  - D. Thursday
- Question 10: Which of the following is NOT correct about demographic information of the individuals recorded in the stops?
  - A. Male counts more than 70% of all the stops.
  - B. White male counts more than 60% of all the white stops.
  - C. White male counts more than 60% of all the male stops.
  - D. Among White, Black, Hispanic and Asian, Hispanic has the highest male to female ratio in terms of number of stops in the data.

In [ ]:
```
### your code here
```

3. Impute

- Fill in missing values in the column `subject_age` using column median.
- Fill in missing values in the column `reason_for_stop` with 'unknown'.
- Replace 'asian/pacific islander' in column `subject_race` with 'asian'.
- Then anwser Question 11 - 15 on Canvas.

- Question 11: There are two columns in the data contains the exact same values. Which are the two columns?
  - A. `contraband_found` and `search_conducted`.
  - B. `search_conducted` and `search_vehicle`.
  - C. `search_vehicle` and `search_basis`.
  - D. `search_basis` and `contraband_found`.
- Question 12: The columns `contraband_found` and `search_basis` have the same number of missing values. What is the best way of handling these missing values?
  - A. Drop all the rows with missing values in these 2 columns.
  - B. Drop these 2 columns.
  - C. Ignore these missing values.
  - D. Drop all the rows with missing values in these 2 columns and also drop these 2 columns.
- Question 13: Which of the following statement is NOT true about search conducted for each race?
  - A. Overall search is conducted in about 6% of all the stops.
  - B. After being stooped, race "Black" are most likely to get searched.
  - D. After being stopped, race "other" are least likely to get searched.
  - C. After being stopped, race "Hispanic" are more than three times more likely to be searched than White.
- Question 14: Which of the following statment is NOT true about search conducted for each race and gender?
  - A. After being stopped, the proportion of male being searched is more than twice of the proportion of female being searched.
  - B. After being stopped, race "Aisan" has the least gender disparity in terms of likelihood of being searched.
  - C. After being stopped, race "Hispanic" has the most gender disparity in terms of likelihood of being searched.
  - D. All of them are True.
- Question 15: What is the `subject_age` in `police_stop` that have the most stops before and after filling in missing values?
  - A. 25 and 25.
  - B. 25 and 30.
  - C. 30 and 35.
  - D. 25 and 35.

In [ ]:
```
### your code here
```

4. Explore

- Anwser Question 16 - 20 on Canvas. For each question, please upload a screen shot of your plot. Make sure to add some labels and a title.

- Question 16: Create a plot to look look at subject_race distribution for all the stops.
- Question 17: Create a plot to look at subject_age distribution for all the stops.
- Question 18: Create a plot to look at distribution of subject_age for different subject_race.
- Question 19: Create a plot to look at total number of stops over the years.
- Question 20: Create a plot to look at the percentage of search conducted for different races and genders.

In [ ]:
```
### your code here
```

5. Test

Let's look at the search conducted percentage by month. The month of August have a higher search conducted percentage. We would like to test if the month of August is like a random sample from the entire data set (consider the entire data set as your population). Anwser Question 21 - 25 on Canvas.

- Question 21: What is the search conducted percentage for the month of August?
  - A. 6.10%
  - B. 6.12%
  - C. 6.14%
  - D. 6.16%
- Question 22: What would be the Null hypothesis?
  - A. The month of August is like a random sample from the entire data set.
  - B. The month of August has the exact same search conducted percentage as the entire data set.
  - C. The month of August has a higher search conducted percentage than the entire data set.
  - D. The month of August has a lower search conducted percentage than the entire data set.
- Question 23: What would be the observed test statistic?
  - A. Overall search conducted percentage.
  - B. August search conducted percentage.
  - C. Overall search conducted percentage except August.
  - D. Randomly sample a subset of the data and calculate the search conducted percentage.
- Question 24: How to generate a simulated test statistic?
  - A. Bootstrap the entire data, and calculate the search conducted percentage using a bootstrap resample.
  - B. Bootstrap the August data, and calculate the search conducted percentage using a bootstrap resample.
  - C. Randomly sample a subset of the August data and calculate the search conducted percentage.
  - D. Randomly sample a subset of the data and calculate the search conducted percentage.
- Question 25: Conduct the hypothesis testing using one-sided alternative: August search conducted percentage is higher than overall search conducted percentage, and generate 1000 simulated stats. Using significance level alpha = 1%, what is the p-value and the conclusion?
  - A. p-value is less than 1%, reject the null.
  - B. p-value is more than 1%, reject the null.
  - C. p-value is less than 1%, fail to reject the null.
  - D. p-value is more than 1%, fail to reject the null.

In [ ]:
```
### your code here
```

6. Model

Fit the following two models and anwser Question 25 - 30 on Canvas.

- Model 1: Logistic regression model to predict search conducted based on demographic. Please make sure you first convert `search_conducted` to False = 0 and True = 1, so that the model will predict the probability of search conducted, not the other probability. Use the following variables as predictors:
  - subject_age
  - subject_race
  - subject_sex
- Model 2: Random forest model to predict the race based on other columns. Make sure you first split data into train and test using 50/50 split, and set `random_state = 2024` . Grow 100 trees and use all default values for other hyperparameters in random forest, and also set `random_state = 2024` . Train your model and make prediction on test set. Use the following variables as predictors:
  - subject_age
  - subject_sex: you will have to convert this to dummy variable or boolean type
  - search_conducted
  - year
  - month
  - day
  - weekday

- Question 26: In model 1 Logistic Regression, which race and gender did the model choose as baseline category?
  - A. White male.
  - B. White female.
  - C. Asian male.
  - D. Asian female.
- Question 27: Based on model 1 Logistic Regression result, which race and gender is most and least likely to be searched given the same age?
  - A. most likely: black male; least likely: asian female.
  - B. most likely: hispanic male; least likely: white female.
  - C. most likely: black male; least likely: white female.
  - D. most likely: hispanic male; least likely: white female.
- Question 28: Based on model 1 Logistic Regression result, are younger people or older people more likely to be searched given the same race and gender?
  - A. Younger people are more likely to get searched.
  - B. Older people are more likely to get searched.
  - C. Both younger and older people are more likely to get searched.
  - D. Can not make any conclusion based on model result.
- Question 29: Based on model 2 Random Forest result, what is the accuracy score of the prediction on train set?
  - A. higher than 80%.
  - B. between 60% to 80%.
  - C. between 40% to 60%.
  - D. lower than 40%.
- Question 30: Based on model 2 Random Forest result, what is the accuracy score of the prediction on test set?
  - A. higher than 80%.
  - B. between 60% to 80%.
  - C. between 40% to 60%.
  - D. lower than 40%.

In [ ]:
```
### your code here
```

Congratulation, you are done with this exam!