CSE 544, Fall 2018, Probability and Statistics for Data Science

Assignment 6: Mini-project Due: 12/07, submit to Amogha in NCS 336, 1:00 pm
(5 hypotheses, 50 points total)

I/We understand and agree to the following:
(a) Academic dishonesty will result in an 'F' grade and referral to the Academic Judiciary.
(b) Late submission, beyond the 'due' date/time, will result in a score of 0 on this assignment.

(write down the name of all collaborating students on the line below)

Datasets and hypotheses: Choose one dataset

1. Traffic violations in USA. https://www.kaggle.com/felix4guti/traffic-violations-in-usa
   a. Time of violations is a uniform distribution (0, 24 hours). Use **KS test** to check this hypothesis. Also check this for the subset of data with Male and Female drivers. Use the smallest granularity of time, minutes to quantize the data. **10 points**
   b. Compare means: Among the cars stopped, Honda and Toyota cars have same amount of chances of getting citation and warning. Use 2-sample Wald's test to compare the means. Consider citation and warning as indicator RVs. **10 points**
   c. Distribution of number of citations as a function of time (0, 24 hours) is the same for Wheaton and Silver Spring districts. Check this hypothesis via Permutation test with # random permutations = 100 and 10,000. Repeat for number of warnings data. **10 points**
2. Craig list trucks data: https://www.kaggle.com/austinreese/craigslist-carstrucks-data
   a. The age of cars available in Craigslist follows normal distribution. Use **KS test** to compare distributions N(10,3), N(10,5), N(12,5) where mean and standard deviations are in years. **10 points**
   b. Black color cars are more valuable than blue color cars. Use Wald's 2-sample test to compare means. **10 points**
   c. Distribution of age of automatic transmissions sold is the same as that of manual transmissions sold. Check via Permutation test with # random permutations = 100 and 10,000. Age is calculated as (2019 - model_year). **10 points**
3. P2P lending data: https://www.kaggle.com/skihikingkevin/online-p2p-lending
   a. Principal paid is exponential distribution. Use **KS test** to compare principal paid with exponential distribution with mean = 500, 1K, 2K. **10 points**
   b. Interest paid and principal paid are drawn from same distribution. Use Permutation test with # random permutations = 100 and 10,000. **10 points**
   c. Borrower interest rate is higher for #installments less than 300. Compare mean interest rate for #installments using Wald's 2-sample test. **10 points**

Note:

1. For KS test, use 0.15 as the criteria for Accept/Reject.
2. For Permutation test, use 0.05 as the p-value threshold for Accept/Reject.
3. For Wald's 2-sample tests, use α=0.05 when comparing with $z_{\alpha/2}$.
4. For **your assigned dataset**, propose and test **two more hypotheses**, each worth **10 points** each.
5. **Deliverables:** Code, summary of each hypothesis test with all required steps. For the two hypotheses that are proposed, a paragraph about why that hypothesis is useful in practice.