Final Exam **CS3300** **Winter 2020–2021**

1. For the given file – cars.csv – create a Jupyter notebook and perform the following operations. Attach the PDF of this notebook with your exam submission. (20 points)

a) Read in a CSV file to a dataframe.

b) Select the columns for Make, MPG, Model, Model Year, and Weight and store in a new variable

c) Add a binary feature for eco-friendlier cars, where eco-friendlier is defined as an MPG over 20 miles per gallon.

d) Convert the column Make to a categorical variable

e) Convert the values in a column relating to gas mileage from miles per gallon to km per gallon by multiplying each element by 1.6 and store them as a new feature.

f) Display the number of null entries and datatypes of the edited dataframe.

g) Make a scatter plot for km vs weight

2. Describe the differences between Numpy arrays and Pandas DataFrames in terms of indexing, number of dimensions, homogeneity or heterogeneity of data types in the same data structure, and use cases. (6 points)

3. An important aspect of data science is understanding what your data set consists of. Use the following table to answer the questions below.

Year	Make	Model	Doors	Cylinders	Color	Type	Sale Date	Sale Price
2015	Toyyota	Rav4	4.0	4	Grey	Crossover	1-23-2016	24977.56
2011	Honda	CRV	4	4	Green	Crossover	27-05-2011	28578.23
2017	Toyota	Prius Prime	4.0	4 (EV)	Gray	Sedan	2017-07-05	29856.97
2019	Tesla	Model Y	4.0	0 (EV)	N/A	Crossover	2019-10-01	71000.67
1919	Jeep	Grand Cherokee Trackhawk	4	8	Blue	SUV	Tuesday, October 1, 2010	93625.23

a. Identify all of the issues with the data set that will need to be fixed in the data cleaning step. Assume implementation in Pandas. (6 points)

b. Infer the appropriate types (Boolean, string, categorical, float, int, date time) for each column from the data given above. Assume you are not creating new variables yet. (9 points)

Question 3 cont.

c. You want to analyze the relationships below. Describe how you would do this. (8 points)

i) Number of cylinders versus sale price

ii) Whether or not a vehicle is electric versus sale price.

d. List the appropriate type of plot for each independent variable versus price. (4 points)

i) Make

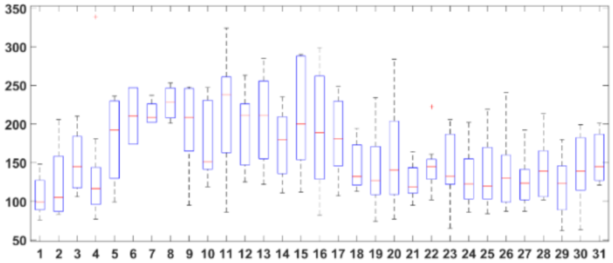
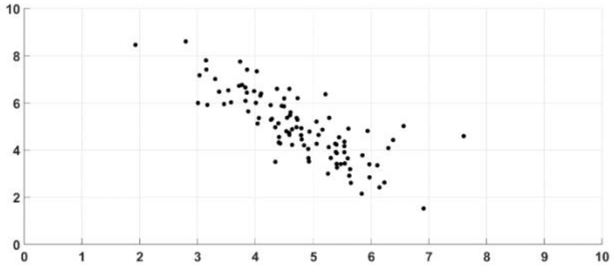
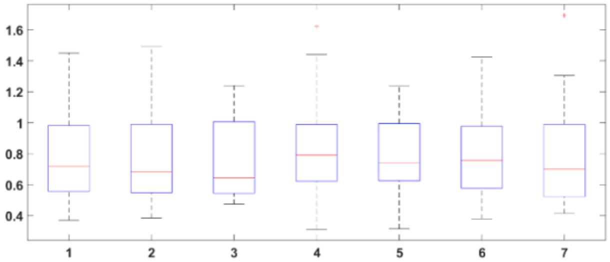
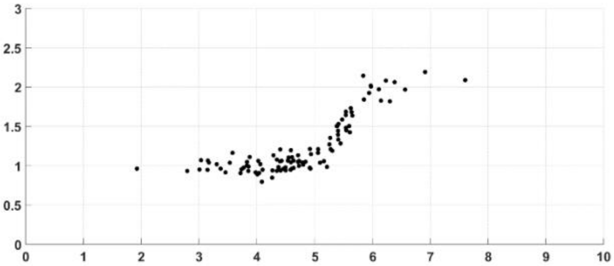
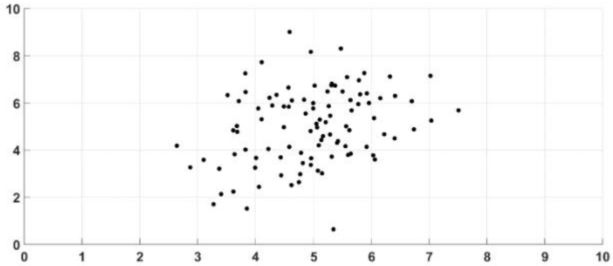
ii) Doors

iii) Type

iv) Sale Date

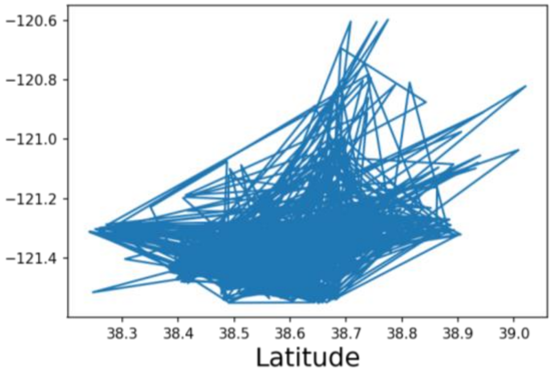
e. Write 2 testable hypotheses (6 points)

4. For the following plots, indicate whether the variables are strongly, somewhat, or weakly correlated. (10 points)

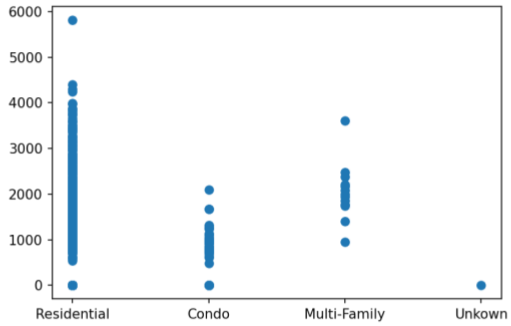
a.		
b.		
c.		
d.		
e.		

5. Your colleague makes several plots from the Sacramento Housing data set. For each plot, determine (i) describe the data types of the two variables used, (ii) whether the plot shown is appropriate for those data types and if not, what type of plot should be used, and (iii) if the plot has all of the necessary elements.

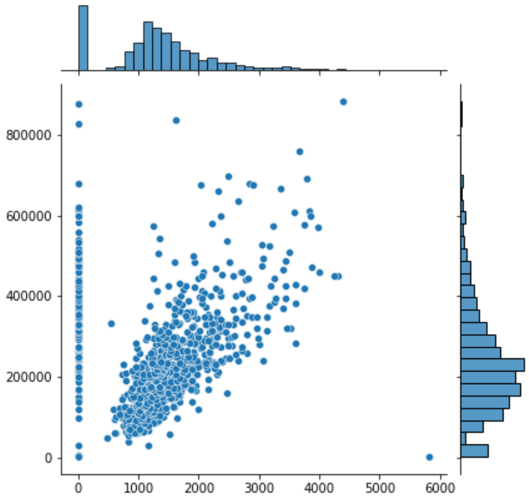
a. Latitude and Longitude (3 points)



b. Square Footage and Property Type (3 points)

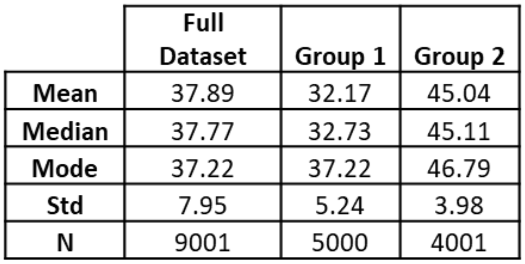


c. Price and Square Footage (3 points)

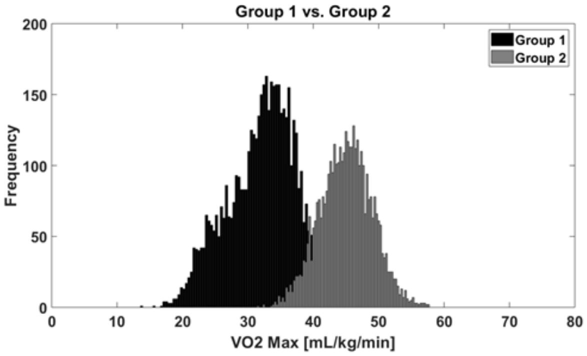


6. You have joined a group of medical researchers to help them process data from their experiment. These scientists are studying oxygen use in endurance athletes and the general population via a measurement called VO2 max. The higher the measurement, the better the individual is able to use oxygen.

You receive a csv file from them with the names of over 9,000 subjects, their individual VO2 max measurements, and a label indicating that they belong in either from group 1 or group 2. Unfortunately, the scientists did not elaborate on what the groups mean. You have done some preliminary processing of the data; calculated some summary statistics; and have plotted the data with and without the labels. Take some time to consider the data and answer the following questions.



	Full Dataset	Group 1	Group 2
Mean	37.89	32.17	45.04
Median	37.77	32.73	45.11
Mode	37.22	37.22	46.79
Std	7.95	5.24	3.98
N	9001	5000	4001



a. Those scientists were total jerks – they didn’t bother to tell you their hypothesis or what the group labels actually mean. Given the data that you have, generate a hypothesis for this dataset. (3 points)

b. It does appear that there is a difference between the groups – calculate or illustrate (using a diagram) the effect size between these populations. (2 points)

c. You use a statistical test to see if there are differences between groups 1 and 2. The statistical test estimates a p-value of 0.068. For a significance level (α) of 0.05, determine if you should accept the null hypothesis or reject it in favor of the alternative hypothesis. (2 points)

Null: there is no difference
Alternative hypothesis: there is a difference

d. What could be done to increase your statistical power? (5 points)

e. As with any appropriate analysis, it is important to question where the data came from and whether there may be some sampling bias in it. Do you have any reason to believe this dataset has bias in it? And if so, can you hypothesize the source of this bias? (5 points)

f. Are there any security or ethical concerns with this data? If so, what are they and what would you do to address them? (5 points)