**Applied Data Science Exam – Sections 1 and 2**

**Update 17 April 2015**

**Instructions:**

For sections 1 and 2, we expect them to take you **four to five hours** to complete.

**You must send back the entire exam within 48 hours.**

You may use any program(s) of your choice (R, Python, Stata, SPSS, Excel, SAS, etc.) to assist you, but please paste your answers into this document. We have a strong preference for use of R and/or Python. You may also look online, use notes, textbooks, software instructions, and any software of your choice. Please indicate at the end of your exam what materials you used.

You may NOT ask anyone else for help with this exam. You may NOT include any content that is not your own. Also, do NOT include any content from previous work, classes, etc. All answers must be original.

Please answer the following questions. Suggested timing for each question is just a guide – you may spend more time with some questions and less time with others.

|  |  |  |
| --- | --- | --- |
| **Problem** | **Recommended Time** | **Value** |
| Section 1 | 150 – 180 minutes | 50% |
| Section 2 | 60 – 90 minutes | 50% |

Please respond to all written responses in this document. In addition, please send the following list of files:

• yourname\_part1\_code.txt (if applicable)

• yourname\_part3\_code.txt

• yourname\_part3\_scores.csv

# Section 1

We will be using the datasets in the attached file:

* Dataset A: individual\_data.csv. Individual-level data on each person’s age, race, gender, voting history, party registration, and education. Each person has a unique person\_id and also a household ID (hh\_id) that indicates the household they belong to.
* Dataset B: household\_data.csv. Household-level demographics. Each household has an income, indicator for whether the house is in an urban or rural area, whether the home is owned by the household and whether the household is in a census tract that is in the top, bottom, or middle third of household income.
* Dataset C: registration\_status.csv. Whether the person is registered as a Democrat (is\_registered\_democrat). Note: 1 = Yes, 0 = Unregistered or in another party, Missing = unknown.

## Part 1

All questions in this section pertain to the individual-level data (A). Please write your answers in this document. Please include your code as **yourname\_part1\_code.txt** (if applicable).

1. Create a variable indicating whether each person is a child or adult (18 and older). Call it adult. What percent of the sample is adults?

Adults make up 76.46 % of the entire sample

1. What is the average number of adults per household? What is the average number of children per household?

The average number of adults per household is 1.61

The average number of children per household is 0.50

1. What percent of adults are college graduates? What percentage of people are college graduates? Do you find the coding of this variable (college graduate) reliable?

The percent of adults who are college graduate is 70.12 %

The percent of people (including children) who are college graduates is 70.07 %

By comparing the variables adults\_college\_graduates and collegegradaute\_percentage, the variable is\_college\_graduate should is not reliable since it is expected that collegegraduate\_percentage < adults\_college\_graduates, but examination of the dataset shows that non-adults were shown to be college graduates, which does not make sense

1. Create a variable that indicates the type of household: “single male head of household,” “single female head of household,” “married couple,” or “unmarried couple.” What percent of households are of each type?

Married couples make up 24.34 %, single female head of households make up 9.56 %, single male head of households make up 29.52 %, and unmarried couples make up 36.58 % of the total households.

## Part 2

Use the categories of households you created in Part 1 (“single male,” “single female,” “married couple,” and “unmarried couple”).

If you were unable to create these categories in question 4 above, you can categorize households by whether they contain couples (married or unmarried) (“couples”) and households that do not contain couples (“singles”).

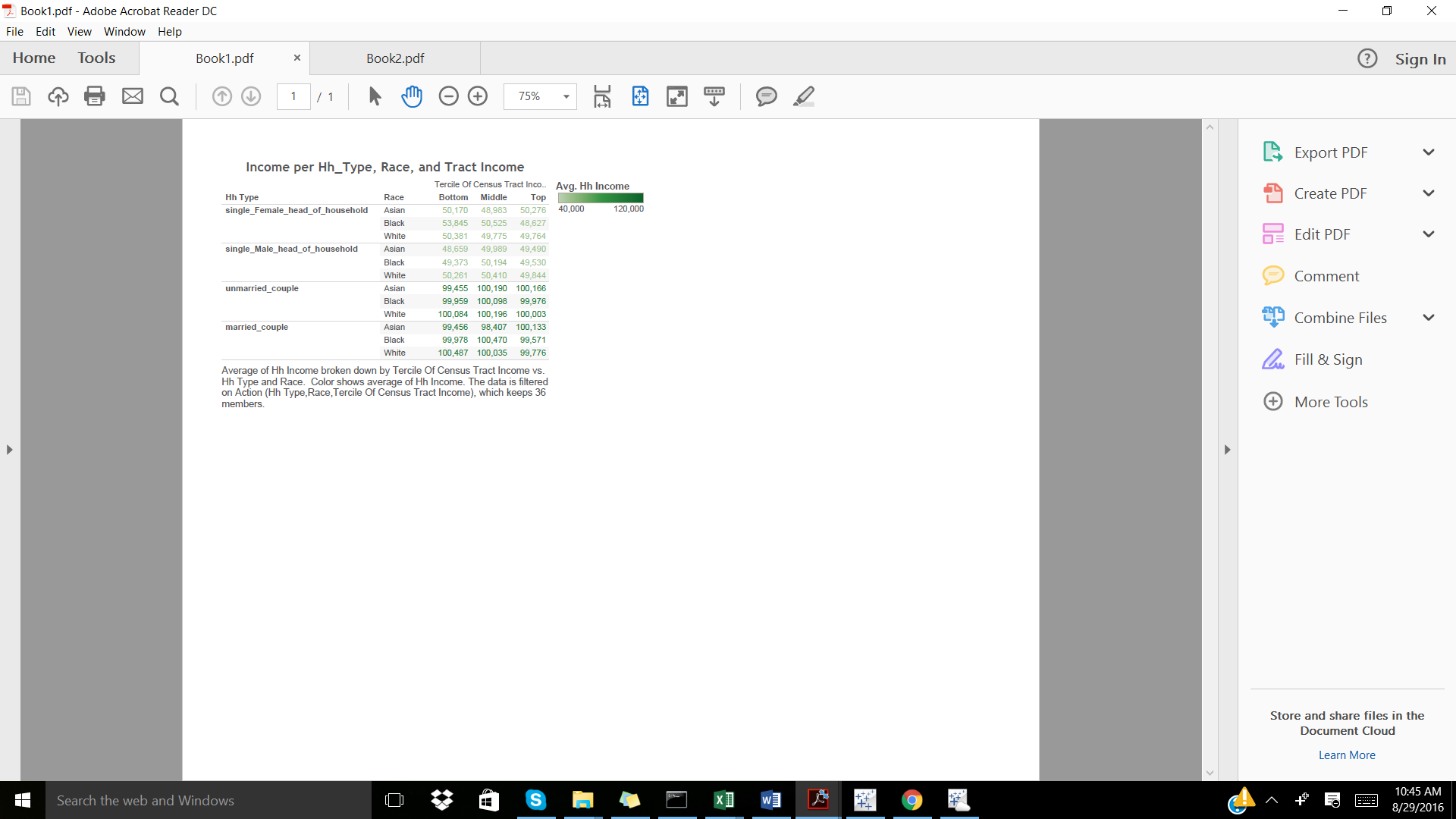
1. Write a 1-2 paragraph comparing households’ socio-economic status across the categories.

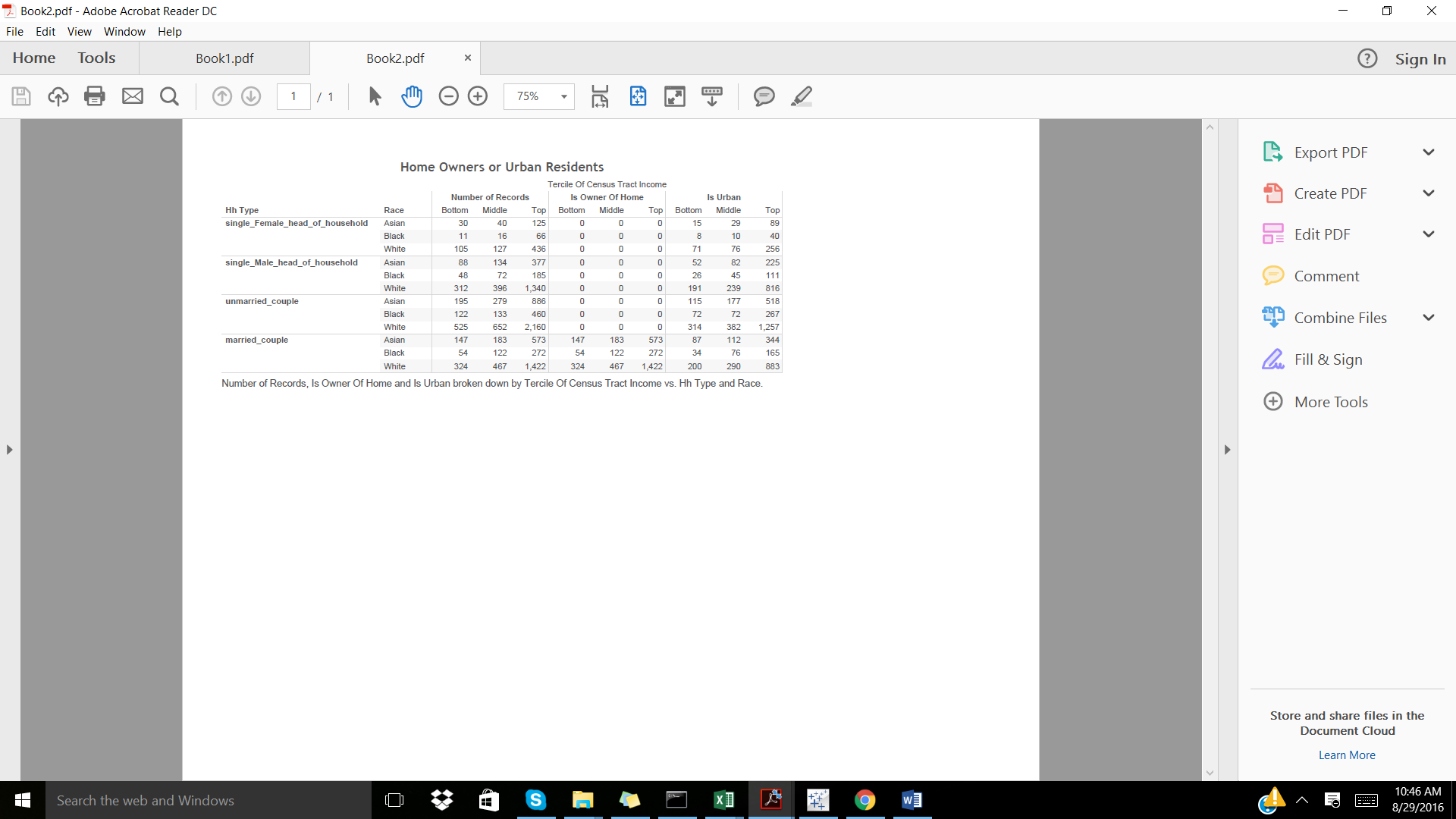
Following analysis of the ‘Income per Hh\_Type, Race, and Tract Income” table, it can be seen that the highest earning group of individuals are white married couples, while black single female head of households earns the least. All home owners are married couple and each household type has residents who reside in urban surroundings.

By using the Tableau Dashboard called “Explore Socio-Economic Information by Household”, it can be seen that among the individuals who are in the Bottom Tercile of Census Tract Income, single male head of households earn the least (independent of race) while married couples earn the most while among the individuals who are in the Top Tercile of Census Tract Income, single male and female head of households earn the least (depending on race) while unmarried couples earn the most, slightly above the earnings of married couples (independent of race). Home owners in both urban and non-urban environments make comparable amount of money while non-home owners who are unmarried couple make the largest sum of money in both urban and non-urban environments.

1. Make a table of summary statistics that illustrate your comparison. Include it below.

Python code was run in order to attain summary statistics for the socio-economic status across the categories (please reference notebook for statistics summary, which is also shown in the Tableau file). To create a more user-friendly visualization and explorer, a Tableau table and dashboard was created.

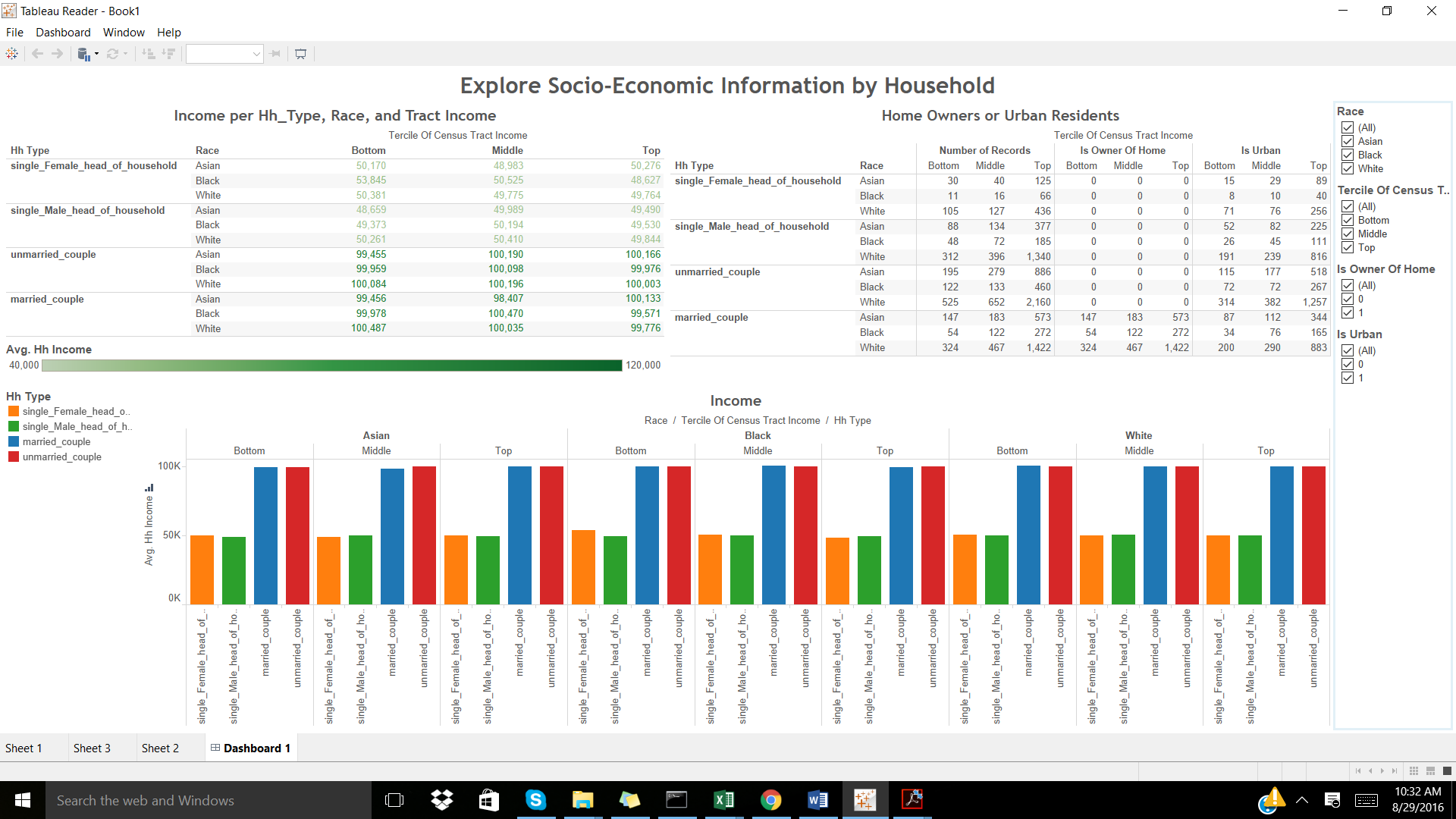




1. Create a client-ready graph to illustrate your comparison. Include it below.

To access the dashboard, please download (for free) the Tableau Reader: <http://www.tableau.com/products/reader>

Open the Book1.twbx and explore. Note: All images, data, etc. can be exported using the Tableau reader.



## Part 3

Please build a model that produces an individual level prediction of a person’s probability to register as a Democrat. In Dataset C, approximately the first 10,000 individuals have known registration status while for the approximately other 10,000 individuals, registration is unknown.

Use Datasets A, B, and/or C to build individual level probability predictions for the people for whom registration is unknown.

You are free to generate any additional variables or to use any that you have already generated for Part 1 and 2. You are free to use any modeling approach.

Please answer the following questions about your approach:

1. List the variables you used and/or considered

The following variables were considered: age, voted\_in\_2012, adult, hh\_income, is\_urban, and is\_owner\_of\_home.

1. Describe the criteria you used to determine how you approached this task and how you developed your model.

I modeled the binary class attribute, is\_registered\_democrat, using a logistic regression with scikit-learn. The above variables were selected as they provide the most insight in regards to the socio-economic status of the individual (as completed in Part 2). The 3 datasets: individual\_editted (with household type included), household\_data, and registration\_status were combined into 1 dataset using the Pandas library. The combined dataset was divided into training (50%) and training data (50%) and the logistic regression was fitted onto the training data first in Python and also in R, producing a model which predicted the class attribute for the testing data set.

1. If given more time, what else would you do?

Through evaluating the accuracy of the model using accuracy\_score, auc\_score, confusion\_matrix, and classification\_report the model developed is in need of stricter feature selection. If I had more time, I would test multiple combinations of variables until I found the ideal combination producing a higher accuracy. I would create categorical data dependency, for the remaining variables as they would improve the model. Furthermore, I would apply 10 fold cross validation when training my training set. The model has poor predictive power, and unfortunately I did not have the time to really dive into the data.

In addition, paste all of your code used in solving this problem into a text file named **yourname\_part3\_code.txt** and send back.

Please send all records that you’ve scored as a CSV file named **yourname\_part3\_scores.csv**. The headers on this csv should be “person\_id” and “score” for person\_id and predicted scores, respectively.

# Section 2

Please answer **three of the six** questions below. Select **one question from each category**.

## Category 1

### Question 1

You and your colleagues have built a model that predicts the likelihood of an individual to not have insurance. Give an example of a client who might be interested in this type of model. How would you explain the meaning and value-add of your model to that client? How would you convince them to incorporate it into their work?

### Question 2

You are advising a company that wants to understand how satisfied their employees are and ways to improve employee satisfaction. What data should they collect and what should they consider while gathering it? Other than collecting data, what else would you do to answer these questions?

Data can be collected from the employees via surveys, these surveys can have distinctive sections for satisfiers and dissatisfiers at the workplace. By rankings of 1 to 10 (10 being the most satisfied), individuals can rate motivators such as recognition, responsibility, nature of work, and personal growth and dissatisfiers such as pay, job security, working conditions, and company management bi-yearly. Along with this satisfaction survey, employee information such as time of employment, department, time with the company, and hierarchical title can provide some insight as to whether or not the individual is overall satisfied.

## Category 2

### Question 1

You and your colleagues have built a model that predicts the likelihood of an individual to not have insurance (higher values = more likely to not have insurance) and sold it to an insurance company. You have been in frequent contact with an analyst at the insurance company, helping her run analysis on the scores. One day she asks you curiously:

*“According to your uninsured model, a certain individual has an uninsured score of 15 (out of 100), but after our company contacted her, it turns out she does not have insurance and hasn’t for a long time. Why has your model assigned her a score of only 15?”*

Without knowing this friend’s individual information, how do you explain this apparent discrepancy to the analyst over an e-mail? **Assume the analyst is a smart, educated person with little or no background in statistics.**

Then suggest a different way that the analyst could confirm the accuracy of the predictive model.

**Answer**

The predictive model developed trained on a specific training data set, which was not provided by the company. Although the predictive model might be effective in predicting the likelihood of an individual to not have insurance, discrepancies lie in the fact that new data was introduced into the system from a new data source. The model must be updated by including the prior testing data along with a subset of the company’s data to train on, by creating a more relevant training data the model will be more accurate, predictive, and helpful in assisting the individual in their analysis. If company specific data was provided prior, I would suggest that although a reasonable model was created to predict the likelihood of an individual to not have insurance, a false positive rate also exists which is associated with the model’s inability to accurately model a subset of data points.

### Question 2

You run a survey (sample size = 2000) for a client that shows 40% of men (size = 1000) prefer dogs to cats while 63% of women (size = 1000) prefer dogs to cats. Is the difference meaningful? How would you explain your results to a smart, educated person with little or no background in statistics?

## Category 3

### Question 1

You and two of your colleagues are working with an online music service that has a “free” and “premium” tier. Giving people free trials to the “premium” tier can be expensive, so they want to target the mostly likely individuals to subscribe. What should the outcome variable be on this model, and what kind of data do you need from the company to determine the **outcome (dependent) variable** only?

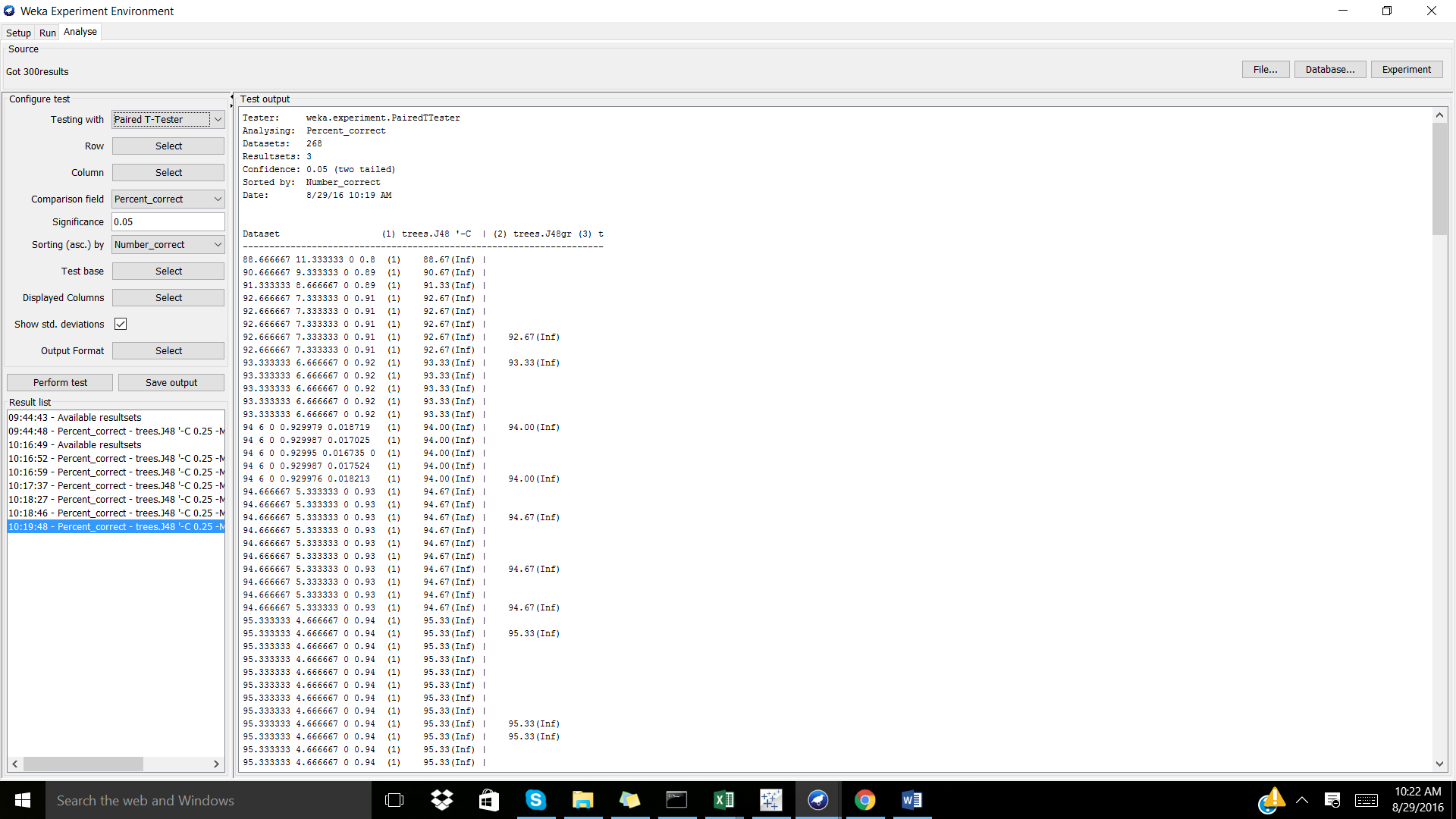
### Question 2

You and two of your colleagues are working with an online music service that has a “free” and “premium” tier. Giving people free trials to the “premium” tier can be expensive, so they want to target the mostly likely individuals to subscribe. You have come up with three different models, using the same set of input variables. How do you evaluate which model performs best?

**Answer**

I would start out by ensuring that the experiment type was the same per model ie. all models used 10 fold cross validation vs. train/test percentage split, by doing so I would ensure that the 3 different models can be compared to one another. In order to create statistically relevant results, each model would be run for an iteration of 10 times at 10 fold cross validation. Paired t-test would be used to determine how well the models performed against one another, with information about the percent correct, percent incorrect, percent unclassified, kappa statistic, mean absolute error, and root mean square error, and standard deviations with sorting by number correct, would be considered. The statistics can be compared between models using Weka/rweka package in R and sklearn.

Screenshot of paired t-test being implemented on 3 different sample models on the same dataset using Experimenter feature in Weka.



# Affirmation

* Please indicate what tools (Websites, textbooks, class notes, software, etc.) you used during the course of this exam and how you used them.

Used: Pandas, numpy, scklearn Python libraries

Used: Tableau software to create dashboard

Used: R

Referenced: Old code listed on my github: https://github.com/OrysyaStus

* Please type your name, location, time and date below to affirm that you neither sought nor received assistance from anyone else during this exam and that all answers are original and entirely your own work and do not include any content created prior to this four hour exam.

“I hereby affirm that I neither sought nor received assistance from anyone else during this exam and that all answers are original and entirely my own work and do not include any content created prior to this four hour exam.”

Orysya Stus. Freiburg, Germany. 5:10 pm (Germany). August 29, 2016.