Homework 1

Due: Thursday 9/27, 11:55pm on Titanium. Prepare your answers as a **single PDF file**.
Group work: You may work in groups of 1-3. Include all group member names in the PDF file. Only one person in the group needs to submit to Titanium.

1. Consider the data collected from the survey conducted during the first class in the attached .csv file. Perform the following steps. Give the corresponding R code (single line), code output, and answer other questions, show plots, if asked. For this question, you may use either base R or the tidyverse library.

a. Load the survey data into a variable called "survey" Hint: use read_csv() (code, output)
b. Which variables are numeric? (code, output)
c. Convert variable Semester to a factor. (code)
d. What is the mean value of Math skills? (code, output)
e. Is the mean skill level in Statistics higher than that in Math? (code, output)
f. What is the mean value of Math skills in Fall semester? (code, output)
g. Is the mean value of Math skills in Fall semester higher than that in Spring? (code, output)
h. How many students have taken CPSC 483?

2. Using the survey data for Q1, create the following plots using ggplot. Give both code and include the plot as an image. Plots can be saved from RStudio or using R commands.

a. A bar graph of variable "CS"
b. The plot above likely has x-axis labels not aligned with the bars. Provide your own breaks to match the variable values/bars. Also, add a plot title. Show both code and paste the plot as an image.
c. Plot (using ggplot) a bar graph of variable "plan_CPSC_483". Show both code and paste the plot as an image.
d. A scatterplot of variables "Math" and "CS"

3. Data wrangling using the tidyverse

Data wrangling cheatsheet:
http://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

Apply the tidyverse's data wrangling verbs to answer these questions about the survey data used in Q1. For each question, **give only the (one line of) code**.

a. List data only for students with skill in CS > 7
b. List data only for students with skill in CS > 7 and skill in Math > 5
c. List data only for students with either skill in CS > 7 or skill in Math > 5
d. List data only for students who are CS majors.
e. Sort data in order of increasing Statistics skill
f. Sort data in order of decreasing Statistics skill
g. Show only the Semester and CS major of students in order of decreasing Statistics skill
h. Add a new variable, Math_Statistics, that indicates the total Math + Statistics skill
i. Show only the Semester and CS major of students in order of decreasing Math_Statistics skill
j. Show only the Semester and CS major of students with 10 highest Statistics skill (Hint: use the min_rank() function which assigns ranks 1, 2, 3, …)
k. Show the average Math_Statistics skill for every Semester.
l. Show the average Math_Statistics skill for every Semester-CS major pair

4. Data reshaping using the tidyverse

a. Consider the attached .csv file "horse_racing.csv" which contains data related to horse racing licensing in New York[1]. The `License` column has two types of values: license numbers and receipt numbers. Load the dataset and transform it such that this column is split into two:
   i. `LicenseOrReceipt`: a factor with two levels "License" and "Receipt"
   ii. `Number`: numeric column with the license/receipt number

   Show (1) your code, and (2) copy & paste the output of the function **`str()`** on your final table.

b. Consider the attached .csv file, "language_diversity.csv," which contains data on the diversity of languages in different countries and other parameters[2].
   a. Is the data "tidy"? Explain your answer in 2-3 sentences.
   b. Convert the data to tidy data. Show (1) your code, and (2) copy & paste the output of the function **`str`** on your final table.

c. Consider the attached .csv file, "diseases.csv," which contains data from Australia on hospitalizations[3].

Diseases	Patientdays_Y2015-16	Separations_Y2015-16	Patientdays_Y2016-17	Separations_Y2016-17

[1] Original dataset: https://data.ny.gov/Government-Finance/Horse-Racing-Licensing/cz9u-yj7m/data
[2] Dataset from: https://github.com/jvcasillas/untidydata#language_diversity
[3] Dataset from: https://www.aihw.gov.au/reports/hospitals/principal-diagnosis-data-cubes/contents/data-cubes

1 Certain infectious and parasitic diseases (A00-B99)	694,007	170,095	771,770	186,034
2 Neoplasms (C00-D48)	2,223,563	666,594	2,235,045	684,075
3 Diseases of the blood and blood−forming organs and certain disorders involving the immune mechanism (D50-D89)	317,085	175,590	335,699	190,568

The first few rows are shown above. Load this file and convert the table to the tidy format shown below. Note the new column names. Show (1) your code, and (2) copy & paste the output of the function **`str`** on your final table. The column names and contents should match (but the order of rows may be different.) (*Hint: this will require multiple transforms from gather/separate/select. Read the file with* ***read_csv****, not read.csv*)

Diseases	Year	Patientdays	Separations
1 Certain infectious and parasitic diseases (A00-B99)	Y2015-16	694,007	170,095
1 Certain infectious and parasitic diseases (A00-B99)	Y2016-17	771,770	186,034
2 Neoplasms (C00-D48)	Y2015-16	2,223,563	666,594
2 Neoplasms (C00-D48)	Y2016-17	2,235,045	684,075

5. Consider this answer posted to Quora.com to “Why is R great for Data Science?” (see attached PDF). **(5 points)**

What are the 5 parts of the R ecosystem?

In your opinion, which of the 5 parts is most important for data science?

Justify your opinion in 2-3 sentences.